REVIEW OF THE STATE OF THE ART OF THE DIGITAL CURATION OF RESEARCH DATA

ALEX BALL

erim1rep091103ab11.pdf

ISSUE DATE: 4th May 2010

Catalogue Entry

Title	Review of the State of the Art of the Digital Curation of Research Data
Creator	Alex Ball (author)
Contributor	Tom Howard
Subject	Open Archival Information Systems; data lifecycles; data deposit; research proposals; data centres; preservation metadata; institutional repositories; cost-benefit analysis
Description	The digital curation of research data is best understood in the context of the data lifecycle, and specifically in the context of data repositories. Disciplinary data centres have established requirements for deposited data, and these requirements are increasingly reflected in requirements and guidance issued by research funding bodies. The digital curation community is active in helping researchers and institutions meet these requirements, producing not only further guidance but a suite of useful standards, technologies and tools. Collectively, these provide a wealth of resources on which the ERIM Project may draw.
Publisher	University of Bath
Date	3rd November 2009 (creation)
Version	1.1
Type	Text
Format	Portable Document Format version 1.4
Resource Identifier	erim1rep091103ab11
Language	English
Rights	

Citation Guidelines

Alex Ball. (2010). *Review of the State of the Art of the Digital Curation of Research Data* (version 1.1). ERIM Project Document erim1rep091103ab11. Bath, UK: University of Bath.

CONTENTS

1 INTRODUCTION

The activity of curating scientific and research data for long-term study is by no means a new one. It is only comparatively recently, however, that this activity has been performed entirely digitally on digital data, and more recently still that it has been taken up outside specialist data centres. It is therefore unsurprising that the curation of digital research data in a generalist context is still an immature discipline. There is however a wealth of experience among specialists from which which generalists can learn, and an active area of research dedicated to making digital curation easier to perform on an institutional and individual level.

The aim of this report is to present the state of the art of the digital curation of research data, in terms of both theoretical understanding and practical application, and note points of particular interest to the ERIM Project. The report begins by reviewing the concepts of data curation and digital curation, and then exploring the terminologies currently in use for describing digital repositories and data lifecycles. Some parallels are also drawn between digital curation practice and design and engineering practice. Existing guidance on data curation from research funders, established data centres and the Digital Curation Centre is summarized in section 3. A review of some important standards and tools that have been developed to assist in research data management and digital repository management is presented in section 4. Finally, a short case study of implementing a new data management plan is presented in section 5, followed by some conclusions and recommendations in section 6.

2 DATA CURATION CONCEPTS

2.1 Data curation and digital curation

The term ‘digital curation’ is relatively new, having been coined in 2001 as the title for a seminar on digital archives, libraries and e-Science [Bea06]. It was given perhaps its most

precise formulation by Lord and Macdonald [LM03, p. 12], who proposed the following definitions for curation, archiving and preservation.

> **Curation.** The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials.
>
> **Archiving.** A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.
>
> **Preservation.** An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

Note that while the definitions are for curation, archiving and preservation *simpliciter*, the authors have digital objects, and specifically datasets, in mind. While these definitions have been influential in the field of digital curation,[1] there are other, looser definitions in circulation. For example, the Digital Curation Centre (DCC) defines digital curation as 'broadly interpreted, ... about maintaining and adding value to a trusted body of digital information for current and future use', with curation in general as 'the active management and appraisal of data over the life-cycle of scholarly and scientific interest' [DCC07].

'Data curation' is an older concept, used to describe the process of selecting, normalizing, annotating and integrating data from journals, reports or third-party databases into a database on a given topic, in order to keep it up-to-date and relevant [Bun+06; Rob03]. Nevertheless, given that in some contexts there is a tendency to treat 'data' and 'digital object' as near synonyms [Gia07], there has in recent years been a similar tendency to treat data curation as a synonym for the much broader concept of digital curation, or perhaps digital curation applied to datasets.

While the priority of the more specialized meaning of 'data curation' is acknowledged, for the remainder of this report the term will not be invested with special significance and will be used to refer to the (general) curation of data, most specifically digital research data. The term 'curation' will be used in the sense outlined by Lord and Macdonald, and therefore will include tasks such as the following.

- Selection of datasets to curate.
- Bit-level preservation of the data.
- Creation, collection and bit-level (or hard-copy) preservation of metadata to support contemporaneous and continuing use of the data: explanatory, technical, contextual, provenance, fixity, and rights information.

1. See, for example, Hitchcock et al. [Hit+05], Bose and Reitsma [BR06], Livingston and Nastasie [LN07], Law, Peng and Demian [LPD05], and the JISC [JIS03].

- Surveillance of the state of practice within the research community, and updating of metadata accordingly.
- Storage of the data and metadata, with levels of security and accessibility appropriate to the content.
- Provision of discovery services for the data; e.g. surfacing descriptive information about the data in local or third-party catalogues, enabling such information to be harvested by arbitrary third-party services.
- Maintenance of linkages with published works, annotation services, and so on; e.g. ensuring data URLs continue to refer correctly, ensuring identifiers remain unique.
- Identification and addition of potential new linkages to emerging data sources.
- Updating of open datasets.
- Provision of transformations/refinements of the data (by hand or automatically) to allow compatibility with previously unsupported workflows, processes and data models.
- Repackaging of data and metadata to allow compatibility with new workflows, processes and (meta)data models.

2.2 Data repository terminology

The preservation of data is an activity that usually takes place within a dedicated repository. As preservation is a high priority activity within curation, an understanding of repositories and their functions is key to producing an effective strategy for curating data.

The Open Archival Information System (OAIS) Reference Model is to date the most influential conceptual framework for describing functions, roles and responsibilities of archival repositories [CCS02]. While the Model was developed with archives of Space Science data in mind, it is framed sufficiently generally to apply to archives holdings all kinds of digital and physical resources, and indeed has become widely adopted in digital repository and digital curation communities.

The Model has two principal aims: to establish a standard terminology for describing the features of archival repositories, and to establish a minimum level of functionality for archival repositories. The Model does not prescribe that the functions or features of the repository should be implemented in any particular way, only that it should be possible to map between the implemented functions and features of the repository, and those described by the Model.

The OAIS Reference Model contains two detailed models. The first is an Information Model which describes the types of objects and information that an OAIS deals with (see Figure 1). The resource to be preserved is the *Data Object*; this concept explicitly covers both *Physical Objects* and *Digital Objects*. The extra information needed to make sense of the Data Object is called *Representation Information*. Representation Information can be structural (e.g. file formats), semantic (e.g. a code book) or some other type (e.g. software); aggregated with a Data Object, it forms *Content Information*, and combined with it, it forms an *Information Object*. It should be noted that the amount of Representation Information needed to make a Data Object understandable to people depends strongly on

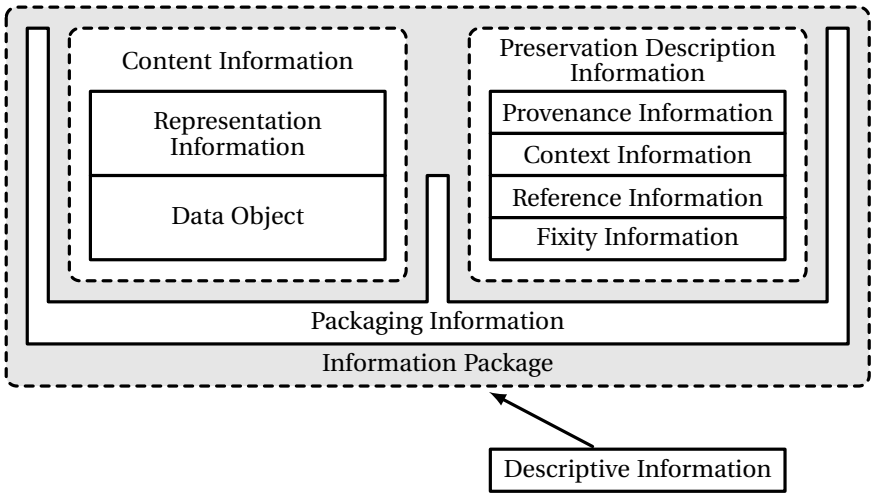


Figure 1: OAIS Information Model [CCS02, pp. 2·5, 4·37].

their pre-existing knowledge and experience: their *Knowledge Base*. It is not feasible for an OAIS to offer sufficient Representation Information to allow everyone to understand the Data Object, so the Reference Model introduces the idea of the *Designated Community*: the community or communities that the OAIS commits itself to supporting. It is the Knowledge Base of Designated Community that determines the amount of Representation Information that the OAIS should be able to supply. As this Knowledge Base evolves over time, one of the functions of the OAIS is to monitor these changes and update the Representation Information held to compensate for them.

The model identifies four further types of information as important for preservation purposes: *Provenance Information* (the source and processing history of the Data Object), *Context Information* (how the Data Object relates to other Data Objects), *Reference Information* (e.g. identifiers) and *Fixity Information* (e.g. checksums). All these pieces of information are listed and related together using *Packaging Information* (e.g. a manifest) to form an *Information Package*. Further *Descriptive Information* (e.g. a catalogue record) should be used to enable the package to be found in a retrieval system. It should be noted that the model does not require or imply that the Information Package must be a literal package – in the digital world, a set of files grouped together within a container file or directory. While there is some administrative benefit to co-locating the information, severe inefficiencies can arise where the same information is relevant for a large number of Data Objects, particularly in the case of Representation Information. For this reason, a number of packaging standards have been developed that permit flexibility in the placement of data: embedded directly in a manifest file, as a separate file within a literal package, or as an external file [ISO05b; Luc05; Pre07; SN07].

The second detailed model is the Functional Model, describing the functions, processes, roles and responsibilities of an OAIS (see Figure 2). While the Model goes into some depth about the processes associated with each functional entity, they may be summarized as follows. Producers submit data and associated metadata to the OAIS in the form of a Submission Information Package (SIP). The *Ingest* entity of the OAIS processes the SIP – performing quality assurance and normalizing the metadata – to produce an Archival Information Package (AIP) and accompanying Descriptive Information. The AIP is trans-

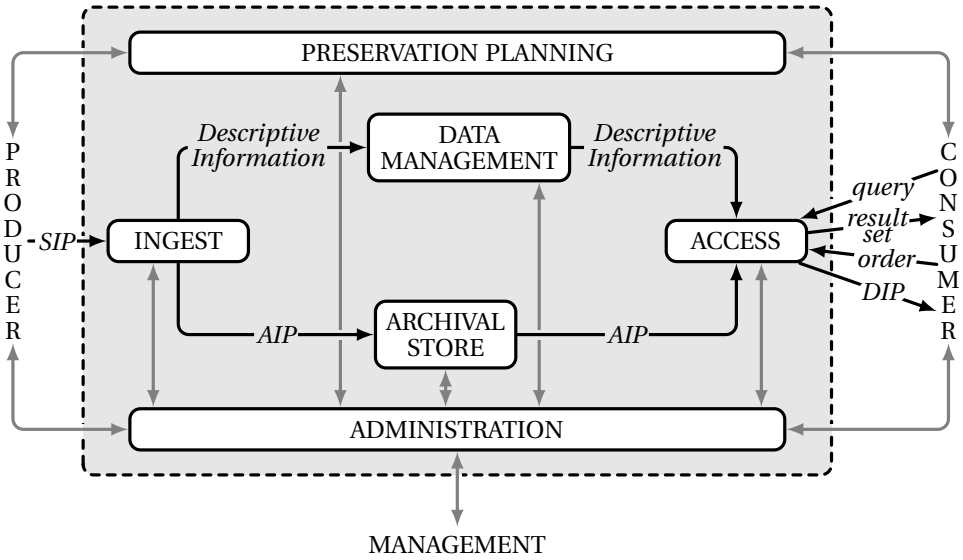


Figure 2: OAIS Functional Model [CCS02, p. 4·1].

ferred to the *Archival Storage* entity, which also performs backups, media refreshment and so on. The Descriptive Information is transferred to the *Data Management* entity, which integrates the information into a catalogue database. When Consumers wish to retrieve a resource from the OAIS, they do so via the *Access* entity of the OAIS. Initially, the Consumer issues a query to the Access entity, which passes it to the Data Management entity. Data Management performs the query and returns a result set, which the Access entity passes back to the Consumer. Alternatively, the Consumer may issue a report request, requiring several queries to be performed; the report is also generated by the Data Management entity before being passed back through the Access entity. Finally, the Consumer places an order through the Access entity; the Access entity retrieves both the Descriptive Information from the Data Management entity and the AIP from the Archival Storage entity, and combines them into a suitable Dissemination Information Package (DIP). The Access entity then delivers the DIP to the Consumer.

The *Administration* entity establishes policies and procedures, negotiates submission agreements, performs audits and manages system configuration. The *Preservation Planning* entity develops preservation strategies, migration plans and packaging designs, and monitors both the community served by the OAIS and the wider technological environment for changes that would affect requirements for Representation Information or Preservation Description Information. Omitted from the diagram are the *Common Services* that underlie the operation of the OAIS: operating system services, network services and security services.

2.2.1 Implications for ERIM

The OAIS Reference Model provides a useful framework for considering how research data ought to be managed. When formulating the ERIM data management plans, the following questions will need to be answered.

- What constitutes the Designated Community for the research data? In other words, of all the communities that may find the data useful or interesting, which should we commit to supporting? This decision would not prevent other communities from using the data, but any additional support needed by these latter communities would have to come from the Designated Community rather than the data archive or repository. Given that the Knowledge Base of the Designated Community has implications for the amount of support a repository needs to give, it is important that the Designated Community is not so broad that support becomes unmanageable, nor so narrow that re-use of the data is unfairly restricted.
- What information would this Designated Community currently need in order to understand and re-use the data? Alternatively, what should a Dissemination Information Package provided to the Designated Community include?
- Which systems and staff will be responsible for delivering OAIS functionality?
- What information does the OAIS need in order to curate the research data? Alternatively, what should an Archival Information Package of the research data contain, and what Descriptive Information would be of most use to Consumers from the Designated Community?
- Which pieces of this information will the OAIS be able to gather through its own processes and from standard sources, and which pieces will need to be supplied by Producers? In other words, what should a Submission Information Package contain?
- How can the workflow of Producers be adjusted to provide the information required for a SIP with minimal effort and disruption?

In order to answer these questions it is also necessary to understand the data lifecycle and in particular the processes by which the data are produced.

2.3 Data lifecycles

As defined in section 2.1, curation involves managing and promoting the use of data from its point of creation. By planning carefully the manner in which data are produced and processed, later curation tasks can be made considerably easier. For example, a repository can be more certain of the preservation actions it can perform if the rights and licensing status of the data has already been clarified, and researchers are more likely to be able to detail the methodologies and workflows they used if they record them at the time.

2.3.1 DCC Curation Lifecycle Model

The DCC has produced a model for aligning curation tasks with the lifecycle stages of a digital object, intended as a planning tool for data creators, curators and users [Hig08]. A graphical representation of the model can be seen as Figure 3.

At the centre of the Model are the digital data, which are here identified with simple and complex digital objects or databases. The Model notes three levels of full-lifecycle actions:

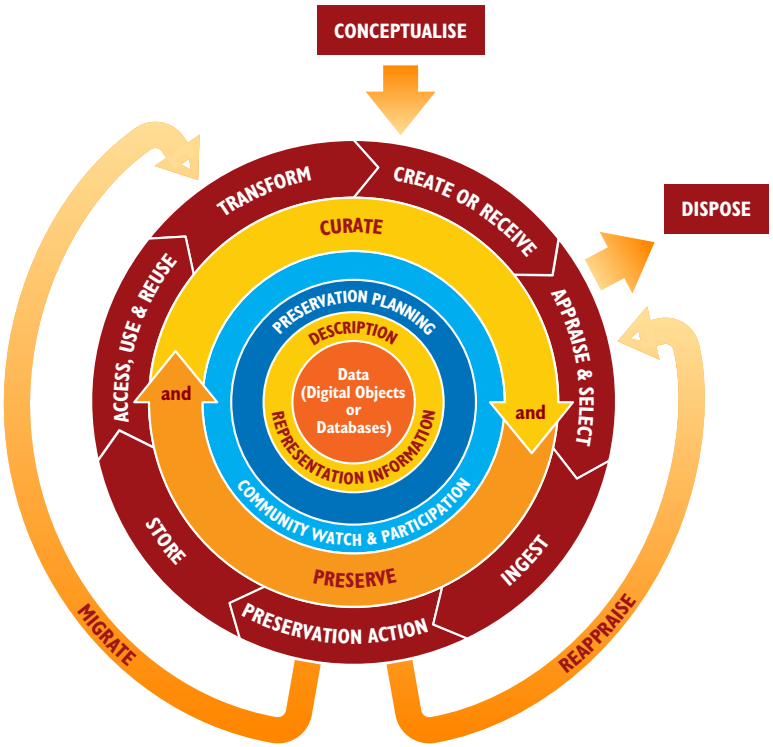


Figure 3: DCC Curation Lifecycle Model.

- *Description* and (management of) *Representation Information*. The creation, collection, preservation and maintenance of sufficient metadata (and recursions thereof) to enable the data to be used and re-used for as long as they have value to justify continued curation.
- *Preservation Planning*. Strategies, policies and procedures for all curation actions.
- *Community Watch and Participation*. The observation of the target community of the data, in order to track changes in their requirements for the data, and participation in the development of standards, tools and software relevant for the data.

The fourth level, *Curate and Preserve*, properly describes most of the actions in the model, but is used here to represent the execution of the planned management and administrative actions supporting curation.

The sequential actions are not exclusively concerned with curation, but rather represent stages of the data lifecycle which ought to have a curation component. They begin with *Conceptualise*: the planning stages of the data generation and collection activities. Aspects such as the capture method will be informed by considerations other than curation – the scientific rigour of the method will be particularly important – but matters such as how the data will be stored, what budget to allocate for curation, and how collection of information important for curation may be automated or otherwise simplified, should be dealt with at this stage.

The curation lifecycle proper begins with the *Create or Receive* stage, where 'create' refers to original data generated and recorded by researchers, and 'receive' refers to pre-existing data collected from other sources. The curation activities at this stage centre on ensuring that all data is accompanied by sufficient administrative, descriptive, structural and technical metadata; ideally, pre-existing data should have these already, but as different researchers and repositories inevitably work to different standards they should be checked for consistency with local policies.

In the next stage, *Appraise and Select*, researchers or data specialists evaluate and select the data to keep for the long term according to documented guidance, policies or legal requirements. Some data may be sent for *Disposal*: this may involve transferring the data to another custodian, although it could mean simple or secure destruction. Again, the nature of the disposal should be driven by documented guidance, policies or legal requirements. The remaining data are sent for *Ingest* by the normal custodian, be that an archive, repository, data centre or some other service. The Ingest stage immediately leads on to the *Preservation Action* stage, which involves an array of different activities: quality control, cataloguing, classifying, generating fixity data, registering semantic and structural metadata, and so on. Any data that fail quality control checks are returned to the originator for further appraisal. This should result either in improvements in the quality of the data (e.g. corrections to data transfer procedures, improved metadata, repackaging of data) and reselection, or disposal. Some data may need to be migrated to a different format, either to normalize it within the system or to reduce risks arising from hardware or obsolescence.

Once the data have completed the *Preservation Action* stage, they pass into *Storage*. This principally refers to the initial committal of the data to storage, but various long-term actions that ensure data remain secure may also be associated with this stage: maintaining the storage hardware, refreshing the media, making backup copies, checking for fixity, and so on.

Once the data have been safely stored, they enter a period of *Access, Use and Re-use*. Curation actions associated with this stage are focussed on keeping the data discoverable and accessible to designated users and re-users. This includes, for example, surfacing descriptive metadata through custom search interfaces or public APIs, and ensuring the preservation metadata held for the data continue to meet the requirements of the designated users and re-users.

Aside from ongoing preservation activities, the story of the archived data considered as an object stops at that point, but several events may cause progression to the *Transform* stage of the lifecycle. A key piece of software or hardware may approach obsolescence, therefore triggering an action to migrate the data to a new format, or a (re-)user may request a subset or other derivation of the data. The end result is a new set of data which starts the lifecycle again. Data created for a repository's internal purposes will, by its nature, pass through the early lifecycle stages rapidly, while data supplied to a (re-)user will progress as normal.

2.3.2 ANDS Data Sharing Verbs

The Australian National Data Service (ANDS) has developed a set of Data Sharing Verbs as a model of the activities that need to be performed in order to make data available to a wider set of users [BT09]. They are used within the service as a structuring technique for operational planning and an advocacy tool for both data producers and data consumers. There are eight verbs in total.

Create. Refers to all the processes involved in creating a data object: generation, collection, aggregation, collation and transformation. During this stage, curation is supported principally by recording appropriate metadata at the earliest opportunity.

Store. The long term preservation of data within an archive, such as an institutional repository, institutional data store or national/international data centres.

Describe. The generation or acquisition of metadata supporting the storage, discovery, access and exploitation of data. These metadata should include descriptions of the people, organizations, and activities that lead to the creation of the data.

Identify. Assigning a persistent identifier to a set of data. ANDS provides a Handle-based identifier service, Identify My Data, for identifying data. Internationally, a project is currently underway to provide a similar system based on DOIs [Ros+09].

Register. Making the existence of data known outside the context of its creation. There are various ways in which this may be achieved: providing a record for the data in a repository catalogue; providing this record in a machine readable format, perhaps using OAI-PMH or RSS, enabling it to be harvested and used in a repository cross-search service; citing the data from a published paper; publishing the data on the Web in RDF format.

Discover. Using a data discovery service to locate data of interest.

Access. Data may be accessed either through an automated system (i.e. using a Web link, perhaps passing through an authentication barrier and/or licensing agreement), or by application to a data custodian.

Exploit. Re-using data. The exploitation of data can only take place if good technical metadata (calibrations, classifications, metrics, etc.) are provided alongside contextual information about the way in which the data were created.

2.3.3 DDI Combined Life Cycle Model

The Data Documentation Initiative (DDI) version 3.0 Conceptual Model [Str04] contains a Combined Life Cycle Model for research data, particularly social science data. The name signifies the fact that the model was formed by combining the initial draft of the model, constructed from a data application perspective, with elements from another model [GK02]. The model is linear for the most part, with one alternative path and one feedback loop; a graphical representation can be found as Figure 4. It has the following eight elements.

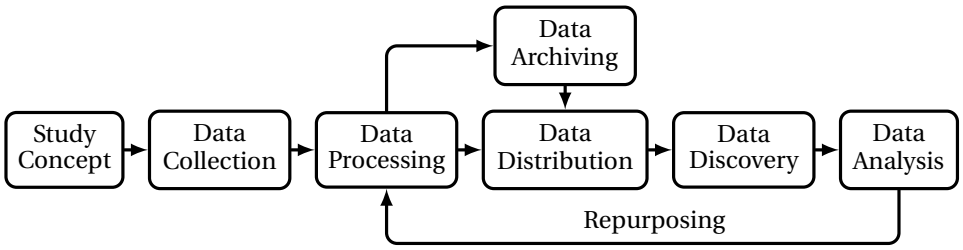


Figure 4: DDI version 3.0 Combined Life Cycle Model.

Study Concept. The earliest stage in the model is the point at which a survey is being designed. The model includes in this not only the choice of research question to answer and the methodology for collecting the requisite data, but also plans for the way in which the data will be processed, analysed and used to answer the question, and the form that answer will take. Researchers should also define at this stage the relationships that will exist between the data products of the research.

Data Collection. Examples given of collection methods and sources include surveys, censuses, voting or health records, commerce statistics or Web-based collections. Primary and secondary data sources should be clearly distinguished.

Data Processing. Once the input data has been assembled, it is processed and analysed to produce output data (e.g. a statistic or set thereof) that answers the research question. These output data may be recorded in a machine-readable form or human-orientated form such as a technical report.

Data Archiving. In order to ensure long-term access to the data, they should be passed to an archive rather than merely kept by researchers. The archive not only preserves the data (and metadata) but also adds value to them over time.

Data Distribution. The data are distributed to users either directly or via a library or data archive.

Data Discovery. The data may be publicised through books, journal publications, Web pages or other online services.

Data Analysis. The data may be used by others within the bounds of the original conceptualization; for example, picking out key statistics for a research report.

Repurposing. The data may also be used within a different conceptual framework; examples include sampling or restructuring the data, combining the data with other similar sets, or producing pedagogic materials.

Within the DDI standard, this model was used to group metadata requirements into five modules: study conception, data collection process, logical structure of encoded data, physical structure of encoded data, and archiving.

2.3.4 UK Data Archive Data Lifecycle

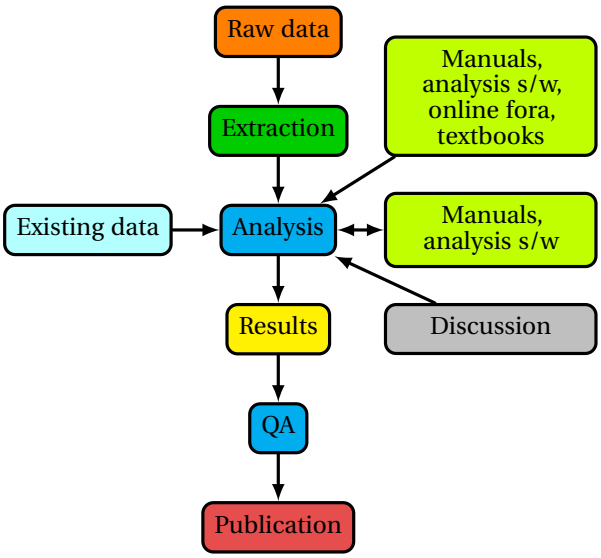
The UK Data Archive provides a data lifecycle model as an aid to researchers considering how data management relates to the lifecycle of a research project [Ukda]. The model defines the following six stages.

1. Data creation
    - research design
    - data management planning
    - data collection (surveying, experimentation, measuring, etc.)
    - data entry or digitization
    - data checking and cleaning
2. Data analysis
    - analysis
    - derived data creation
    - creation of data documentation
3. End of research
    - research outputs
    - preparing data for preservation
4. Preservation of data
    - storage of data
    - migration to suitable format/medium
    - metadata creation
5. Distribution/publication of data
6. Re-use of data
    - by same researcher
    - by other researchers

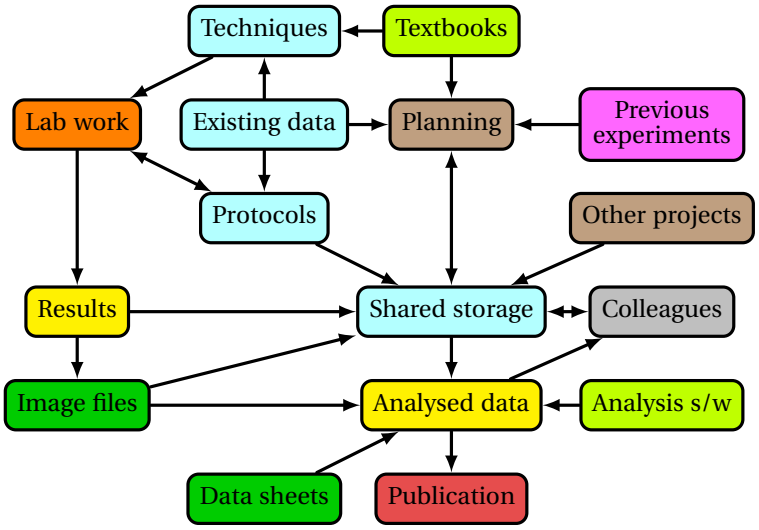
2.3.5 Detailed data flows

The data lifecycle models described above present only a simple view of the lifecycle of data, approximately following a pattern of data generation, collection and recording, followed by a processing stage in which various forms of data transformations and analysis take place, followed by either disposal, archiving or direct dissemination (assuming the data are not simply neglected). The archiving stage permits further dissemination, of course, which in turn allows for the re-use and re-purposing of data. This view hides a great deal of the complexity that exists in real-world research; while the OAIS and DCC Curation Lifecycle models flesh out the detail of the archival stage in the data lifecycle, neither these nor the other models above provide a similar amount of detail when modelling the early stages of the lifecycle. This is not entirely surprising given that the activities that go on in these stages vary enormously between and within disciplines, and are rarely documented in depth.

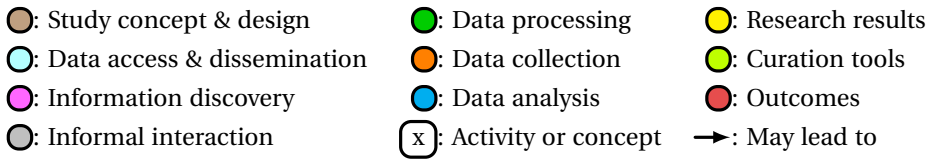
This point is well illustrated by a study conducted by the Research Information Network (RIN) in which a series of case studies were performed examining patterns of information flow, use and re-use within the life sciences [WP09]. The studies found divergence in practice, between life scientists in the same team, between teams in the same discipline, and between disciplines; similarly, there was a divergence between actual practice and



(a) Animal genetics and animal disease genetics



(b) Regenerative medicine



(c) Legend

Figure 5: Simplified information flow maps based on RIN case studies.

the recommendations of policy-makers and information service providers. Part of the research process for these studies involved the use of 'probes': five-day diaries in the style of lab books, in which researchers recorded the information they had used (and whence it had come, what it was used for) and the information they had created (and how and with whom it would be shared). An information flow map was derived from each probe; these maps were then simplified and aggregated to give an information flow map for each case study. Versions of two of these eight aggregated maps may be seen in Figure 5, further simplified to show the gross structure of the information flows.

It may be seen even with these two simplified maps that a research process is a complicated and highly specific array of activities and information flows, and the study concludes that without performing vastly more case studies of this type it would not be possible to deduce a common underlying model. Even if such a model could be devised, the probes indicate that there are real differences in emphasis between disciplines: note that in the Regenerative Medicine map, for example, data analysis clearly takes place but does not appear as a distinct task.

One of the criticisms that could be levelled at the RIN study is that the approach used to map the information flows was too easily coloured by the emphases placed on particular operations by different researchers, and therefore it is unsurprising that common patterns failed to emerge. The flow maps themselves are poorly defined, with nodes standing for either activities (extraction, analysis) or concepts (storage, software, textbooks) and edges indicating a loose progression. A more rigorous approach to process modelling, such as IDEF3 [May+95] or UML [OMG09] activity diagrams, may have provided better data from which to derive a model.

One of the other advantages of using a more rigorous system to document workflows is that it becomes possible to communicate the workflow so precisely that it can be fed into an automated system and rerun automatically. As an example of this, myExperiment is a virtual research environment and social network for sharing re-usable (executable) data processing workflows.[2] By the end of 2009, over 870 workflows had been contributed, primarily from the field of bioinformatics, between them using 25 different formats [DGS09]. The majority of these workflows are written for use with Taverna Workbench,[3] specialist software for running a series of automated operations on data, and range from relatively straightforward to highly complex; an example workflow can be found as Figure 6.

2.4 Data derivation terminology

As part of the development of the Earth Observing System (EOS), the National Aeronautics and Space Administration (NASA) developed terminology for describing the extent to which instrument data has been processed. These data processing levels are defined as follows [Ear86, p. 5].

> **Level 0.** Reconstructed unprocessed instrument data at full resolutions.

2. MyExperiment Web site, URL: `http://www.myexperiment.org/`
3. Taverna Workbench Web site, URL: `http://www.taverna.org.uk/`

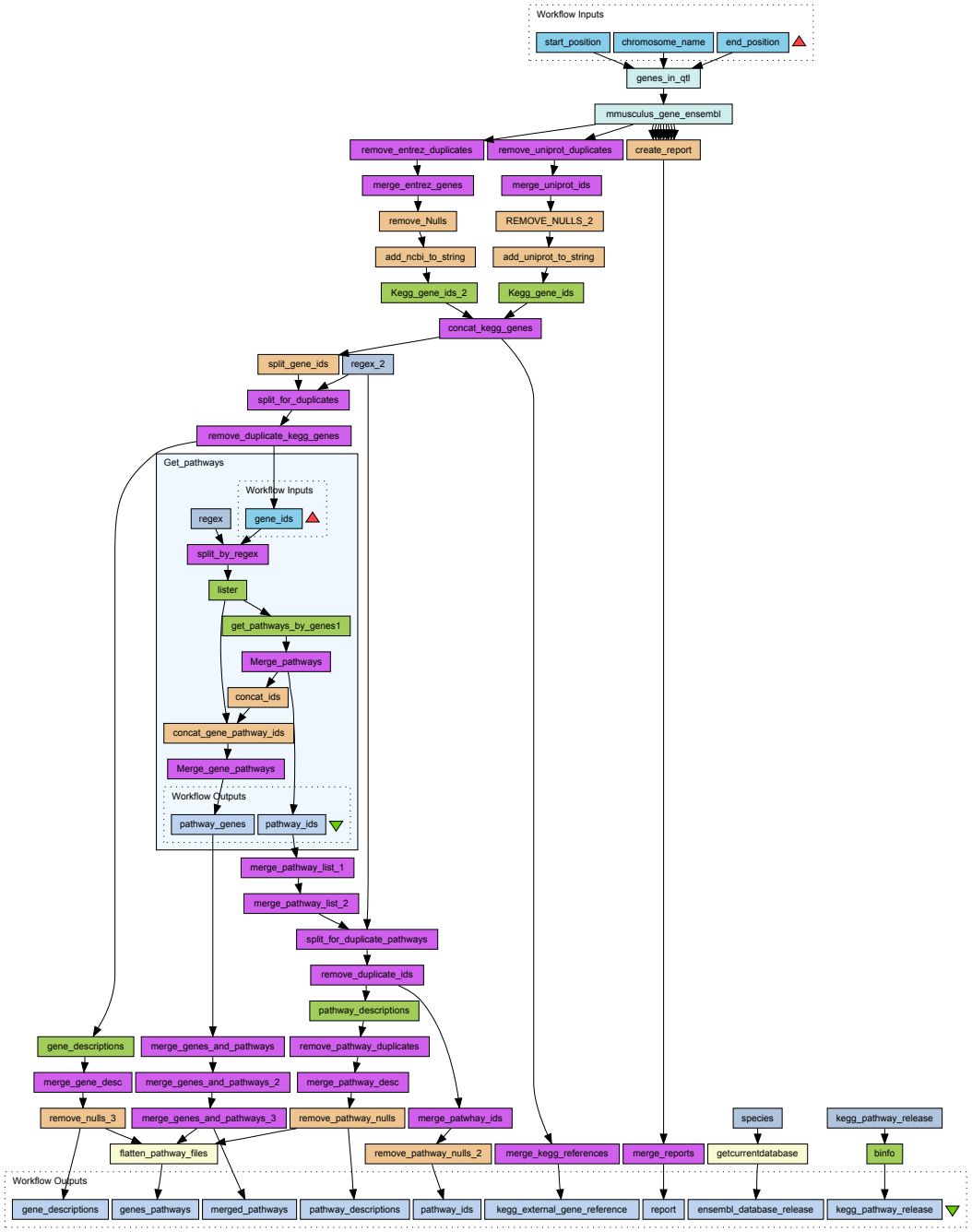


Figure 6: Example research data workflow, ⓒⓘⓞ Fisher [Fis09].

Level 1A. Reconstructed unprocessed instrument data at full resolution, time referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters (i.e., platform ephemeris) computed and appended but not applied to the Level 0 data.

Level 1B. Level 1A data that has been processed to sensor units (i.e., radar backscatter cross section, brightness temperature, etc.). Not all instruments will have a Level 1B equivalent.

Level 2. Derived environmental variables (e.g., ocean wave height, soil moisture, ice concentration) at the same resolution and location as the Level 1 source data.

Level 3. Variables mapped on uniform spacetime grid scales, usually with some completeness and consistency properties (e.g., missing points interpolated, complete regions mosaicked together from multiple orbits).

Level 4. Model output or results from analyses of lower-level data (i.e., variables that were not measured by the instruments but instead are derived from these measurements).

Under this scheme data do not have significant scientific utility until they reach at least Level 1, but most scientific applications require data processed to at least Level 2, and therefore this is the level that has most long-standing usefulness. Level 3 datasets tend to be much smaller than Level 2 datasets, and are generally useful for many applications; they are also much easier to combine with other similar datasets.

A slightly different scheme of levels was devised by the Space Science Board Committee on Data Management and Computation (CODMAC): see Table 1.

2.5 Parallels with design and engineering practice

There is a long history within Engineering of using standards to overcome common challenges, to extend the reach of individual products and to remove barriers to trade. The International Telegraph Union was set up in 1865 to ensure the global interoperability of telegraphy equipment. In 1884, the American Institute of Electrical Engineers was set up and quickly began producing national electrical standards. The American National Standards Institute (ANSI) began life as the American Engineering Standards Committee (AESC), formed from five professional engineering societies and three government agencies [Mig07, pp. 28·2–28·6].

Another early example of standardization is the gauge of railway tracks [Hil92]. In the UK, there was initially some diversity in rail gauge, with the Liverpool and Manchester Railway adopting a gauge of 1435 mm, the Eastern Counties Railway a gauge of 1524 mm, and the Great Western Railway a much broader gauge of 2140 mm. As the need for services to cross railway boundaries became apparent, the break of gauge became a significant issue and in 1846, the gauge of new railways was standardized by Act of Parliament to 1435 mm. By 1892 all main-line railways used this standard gauge. In the US, there was a similar experience, with the movement of grain across railroad boundaries providing

Table 1: Definitions of Space Science Data Levels and Types [Com86, pp. 31–32].

Data Level or Type		Definition	Utility
1	Raw data	Telemetry data with data embedded	Little use to most of science community, except for radio sciences
2	Edited data	Corrected for telemetry errors and split or decommutated into a data set for a given instrument. Sometimes called Experimental Data Record. Data are also tagged with time and location of acquisition. Corresponds to NASA Level 0 data.	Wide use, especially for researchers familiar with instrumentation.
3	Calibrated data	Edited data that are still in units produced by instrument, but that have been corrected so that values are expressed in or are proportional to some physical unit such as radiance. No resampling, so edited data can be reconstructed. NASA Level 1A.	
4	Resampled data	Data that have been resampled in the time or space domains in such as way that the original edited data cannot be reconstructed. Could be calibrated in addition to being resampled. NASA Level 1B.	Wide use, especially for secondary users.
5	Derived data	Derived results, as maps, reports, graphs, etc. NASA Levels 2 through 5.	General way in which information is transferred.
6	Ancillary data	Nonscience data needed to generate calibrated or resampled data sets. Consists of instrument gains, offsets, pointing information for scan platforms, etc.	Needed to be able to convert edited data to calibrated, resampled, or derived data sets.
7	Correlative data	Other science data needed to interpret spaceborne data sets. May include ground-based data observations such as soil type or ocean buoy measurements of wind drift.	Crucial data in many cases to provide ground truth calibration for spaceborne data.
8	User description	Description of why the data were acquired, any peculiarities associated with the data sets, and enough documentation to allow secondary user to extract information from the data.	Important aspect associated with the data that will be even more important for facility-class instruments and for secondary users of data.

Note: We define a secondary user as a researcher not involved with instrumentation design, development, or data acquisition. A secondary user would normally go to a data archive to obtain the required data set.

a significant driver for standardization. By 1886, all North American railroads used the standard gauge of 1435 mm.

More recently ISO 10303 [ISO], otherwise known as the STEP standard, has done much to promote interoperability within engineering by means of application-specific data and information modelling based on a set of generic methods and integrated resources. One of the most successful parts of the standard is the application protocol (AP) that defines a data model and format for Computer-Aided Design (CAD) geometry [ISO05a], providing a vendor-neutral way of exchanging data between CAD systems. As well as expanding to cover more use cases throughout the Engineering domain, STEP is also under constant revision to keep pace with the state of the art. For example, AP 224 [ISO06] came into being to support feature-based CAD modelling, an innovation not catered for by AP 203. Similarly, in the field of computer numerical control (CNC), vendors have been adding incompatible, proprietary extensions to ISO 6983-1 [ISO82] to support newer capabilities of manufacturing machinery. STEP AP 238 [ISO07] was written to remove the need for these extensions by integrating ISO 14649 [ISO03a] and AP 224 data, which could in principle be interpreted directly by a CNC controller. In practice, due the high degree of sophistication this requires, no commercially available controllers are currently able to process AP 238 files, leading to a number of proposed intermediate solutions. One of these involves storing complete AP 238 data structures in a data warehouse, and providing interfaces allowing CAD, Computer-Aided Manufacturing and CNC systems to interact with the AP 238 data [Nas+08]. This approach has two major consequences: (*a*) interoperation is achieved by mapping each tool's native data structure to AP 238, so changing a single tool involves writing just one interface, instead of re-writing and optimizing the entire process chain, and (*b*) from a curatorial perspective, data loss across the interfaces is managed by ensuring all significant properties are catered for in the data warehouse copy, with any additional data on the tool side having the status of possibly ephemeral, convenience data.

The computer industry provides several examples of successful standards, of which perhaps the most successful in the Internet Protocol Suite. As the name suggests, this is the suite of protocols that define how computers interact with one another, of which the most important are the Transmission Control Protocol (TCP) and the Internet Protocol (IP). It is the common support of these protocols by computers of all descriptions that allowed the growth of the Internet in general and rapid growth of the World Wide Web in particular. Again, interoperability is achieved by implementing a small number of standard protocols instead of a different protocol for every other type of device, and there is clear separation between the data needed for Internet communication and the data needed internally by each computer.

Standards are recognized as an important way of reducing the proliferation of sizes, ratings and component-whole interfaces [Tur03, p. 49·4]. This standardization has direct economic advantage in that, for example, it reduces the quantity of spares that have to be stocked (as the same spare can be used in many products), and makes it considerably easier to design a system incorporating many diverse components in multiple configurations. This philosophy is becoming more fashionable in software design, with the rise of service-orientated and process-orientated architecture. It is recommended that the ERIM Project consider how the same philosophy may be applied to the collection and recording of research data.

3 GUIDANCE ON DATA CURATION

As computers have grown more powerful, and processing data has become cheaper, many new possibilities have opened up for re-using existing data, from third-party verification of results to interdisciplinary analysis of diverse datasets. As a result of this and momentum from the Open Access movement, funders of research are becoming increasingly insistent that researchers work up their data into a reusable form from which future research may benefit. In some disciplines there is already a long history of building collections of high-quality data; the good practice and high standards from these disciplines are becoming increasingly refined in the face of ever larger quantities of digital data, and spreading into disciplines with a much shorter history of digital curation. Much of this good practice is also picked up by the digital curation community in general.

This section reviews in turn the requirements placed on researchers and the guidance offered with regard to curating research data, by research funders, UK disciplinary data centres and the Digital Curation Centre.

3.1 Requirements and guidance from funders

3.1.1 Arts and Humanities Research Council

The Arts and Humanities Research Council (AHRC) requires grant holders to make available in an accessible depository any significant electronic resources or datasets generated by the funded research [AHR09]. These resources or datasets must remain available for at least three years after the end of the grant. Archaeological projects must deposit their electronic resources or datasets with the Archaeology Data Service (ADS) within three months of the end of the grant, or else obtain a waiver of deposit from the ADS.

Bidders for funding must complete a Technical Appendix if their project will create an electronic resource as a significant product or by-product. The Technical Appendix is used to demonstrate the technical feasibility of the project, especially the ability of the project to deliver the proposed outputs to the target audiences. It is implicit that the Technical Appendix should demonstrate that the project is capable of meeting the data accessibility requirements above. James and Dunning [JD08] provide the following advice for completing the Technical Appendix.

- *Data development.* This section covers content selection, data and file formats, resource documentation, and guidance received from elsewhere. It should describe how the digital resources arising from the proposed project will be created, organized, delivered and documented, and the methodology to be used when creating software. The role of standards and best practice within the project in the areas of digitization, resource discovery metadata, documentation, access and accessibility, and long-term preservation should be explained.
- *Infrastructural Support.* This section asks for the hardware, software and technical expertise currently available to the bidders, and what additional IT support may be needed. Plans for safeguarding digital assets from loss or corruption should be described in detail.

- *Data preservation.* This section asks for evidence that advice on preserving the project's data or software resources has been sought, and for plans to make these resources sustainable and accessible from a repository. Sustainability is a particular issue if the resources will be hosted in-house, given that the grant money will not be available by that stage.
- *Access.* This section is concerned with how the resource will be disseminated, e.g. by download from a data centre, by download from an institutional Web site, or on optical media. Bidders should go into detail about the technologies to be used, the staff costs of providing the service, whether access will be controlled through user registrations, and so on.
- *Copyright and intellectual property issues.* This sections asks for evidence that advice has been sought on copyright issues. Details should be given of any such issues affecting data or software resources, what has been done to resolve them and the final conclusions of these actions.

3.1.2 Biotechnology and Biological Sciences Research Council

The Biotechnology and Biological Sciences Research Council (BBSRC) expects, but does not make it an absolute requirement, that grant holders should make available the data generated by funded research in a 'timely and responsible manner' for others in the scientific community to re-use [BBS07]. The data should make use of existing standards where available and retained for ten years from the completion of the project. The data should be available from the point at which the associated findings are published, if not earlier; a period of up to three years between generation and release is considered reasonable.

The BBSRC does not prescribe particular approaches, recognizing variance of practice within the fields it funds, but does emphasize data quality, provenance tracking, documentation, citation, and compliance with regulatory requirements. It stresses the importance of using standard formats, methodologies and metadata schemes, noting how this improves the interoperability of the data.

Bidders must provide a statement on data sharing, either explaining why data from the project may not be shared, or outlining a data sharing plan. Suggestions for the content of such a plan include the expected volume, type and content of data; the methodological, metadata and data standards that will be employed and why; any relationships with other datasets; expected secondary uses for the data; anticipated data sharing mechanisms; restrictions that may be placed on the data; final formats for data sharing; and the planned timescale for releasing the data for sharing. The BBSRC encourages researchers to prepare their data for sharing during the period of the grant; nevertheless, long-term data-sharing costs may be included in the full economic costing of bids.

3.1.3 Engineering and Physical Sciences Research Council

The Engineering and Physical Sciences Research Council (EPSRC) does not explicitly encourage the re-use of research data. Rather it strongly recommends storing the data

securely 'for an appropriate time in a durable form under the control of the institution of origin' [EPS06], primarily as protection against allegations of fraud and to protect intellectual property rights.

3.1.4 Economic and Social Research Council

The Economic and Social Research Council (ESRC) requires that data from funded research should be prepared for sharing and, within three months of the end of the grant, offered to the Economic and Social Data Service (ESDS) for deposit within the UK Data Archive (UKDA) or the network of archives co-ordinated by the Qualitative Data Archival Resource Centre (Qualidata) [ESR00]. Any exemptions from these requirements must be applied for, and are reviewed on a case-by-case basis.

The ESRC's attitude to sharing data is explicitly driven by economic efficiency. It prefers researchers to generate new research data rather than to buy existing data, and warns against repeating earlier data generation exercises or purchasing data to which researchers already have licensed access. On the other hand, it promotes the use of important archived datasets, and is not averse to toll-access publication of data, including cases where data are free to UK academia but chargeable elsewhere. It encourages the citation of datasets.

The online submission system for bids requires bidders to answer the following five questions [Esd].

> - If the research involves data collection or acquisition, please indicate how existing datasets have been reviewed and state why currently available datasets are inadequate for this proposed research.
> - Will the research proposed in this application produce new datasets?
> - It is a requirement to offer data for archiving. If you envisage any difficulties in making data available for secondary research, please outline the difficulties.
> - Who are likely to be the potential users of the dataset?
> - Please outline the plans for and cost of preparing and documenting data for archiving to the standards required by the ESDS.

The ESDS provides a list of activities that might be involved in preparing data for archiving, and would therefore need to be costed. These include

- providing data files with a consistent naming and labelling scheme;
- ensuring sensible variable, value and code labels are used within quantitative data tables;
- providing inventories for qualitative data;
- producing transcripts or summaries of interviews and group discussions;
- translating foreign-language transcripts into English if necessary;
- digitizing non-digital sources;

- anonymizing confidential or sensitive data;
- documenting data, research methods, research instruments, etc.

3.1.5 Medical Research Council

The Medical Research Council (MRC) takes a similar view to the BBSRC: it expects, but does not absolutely require, that grant holders should make available the data generated by funded research in a 'timely and responsible manner' for others in the scientific community to re-use [MRC08]. Such re-use should add significant value to the data and fully acknowledge the contribution of the original data creators.

Funding proposals must include a section entitled 'Data sharing and preservation strategy', or else provide cogent arguments for not sharing the data. The strategy should be succinct and summarize the types the data to be generated, expected secondary uses for the data, and plans for preparing and documenting the data ready for sharing and preservation. The expected level of detail for this strategy depends on the nature, scale and cost of the data. Proposals to extend existing datasets also need to what distinctive benefits this would bring, and how sharing the data would provide opportunities for coordination or collaboration.

The MRC also provides some specific guidance on how research data should be managed [MRC05]:

- Personal information must be encoded and anonymized as soon as possible.
- Data should be stored safely and monitored for completeness and accuracy. Raw data should be retained alongside cleaned data.
- Primary data should be retained for at least ten years from project completion. Research records from clinical and public health studies should be retained for twenty years.
- Where notebooks are used, they should be citable (e.g. numbered pages), kept up to date, and reviewed for quality. Consent forms should be kept with the data to which they refer.
- Digital data should be backed up; where possible, important data should be backed up in print form.
- The software used to process the data should be archived as well.
- It should be possible to perform a complete retrospective audit of the data.

3.1.6 Natural Environment Research Council

The Natural Environment Research Council (NERC) provides extensive guidance on data management [NER02]. At the early planning stages, it recommends that researchers answer the following questions:

- What existing datasets will be needed by the project and what, if anything, will it cost to acquire them?
- What datasets will be produced by the project and who will be responsible for their initial management?
- What are the requirements for inter-operability between such datasets and how are they to be met in practice?
- What data (and associated software) standards and quality assurance arrangements should be set in place?
- Is the implication that participants will require a project-specific data service; if so how will it be provided and what will it cost?
- How can the data to be assembled be best exploited, whether scientifically or commercially? Who will take the lead in this exploitation and how will any commercial benefits be shared?
- Should specific data products be published or otherwise promulgated as a direct output from the project?
- Which, if any, of the datasets should be accorded long-term stewardship and become part of NERC's environmental data resource? Reasons should be given for *not* according individual datasets long term stewardship.
- Who will undertake this long term stewardship?
- Should any specific data services set up during the project be perpetuated after it has finished? If so, will continuing funding be required for them, and can a source be identified?
- What are the overall resource implications of the above plans?
- What are the technological implications of the above plans?

Prior to deposit, those holding the data should be cautious about distributing data in such a way as to undermine later commercial exploitation, and should take precautions against damage or loss. When preparing the data for deposit, researchers should use generic processing techniques where possible to ensure compatibility with other, similar datasets. They should process and document the entire dataset, rather than just the data relied on in research papers, and offer 'worked up' data (calibrated, quality controlled) rather than raw.

Data resulting from NERC-funded research must be offered for deposit at the most appropriate NERC designated data centre; there is no default time limit for this to occur, but researchers should be liaising with that data centre throughout the project to establish particular deposit requirements. If data are to become part of NERC's long-term data holdings, they should at a minimum have clear ownership, a full catalogue record (as defined by the data centre), sufficient documentation for independent third-party use, a plan for storage, management and access at a technical/technological level, costings for long-term curation, and a review schedule for determining when the cost of maintaining the data is no longer justified.

3.1.7 Wellcome Trust

The Wellcome Trust expects the data it funds to be as openly available as possible; it also expects those using the data to give due credit to the originators of the data and abide by any terms or conditions imposed [Wel07a; Wel07b]. The Trust provides sustained funding for key data resources.

Bidders must consider how the data generated by the proposed research will be managed and shared. If the research will generate a large quantity of data that could be shared or a data resource for the community, the bid must include a data management and sharing plan. This plan must propose mechanisms by which the data will be shared and specify timescales for this to happen. Embargo periods or access restrictions must be justified. Other issues that may be covered by the plan include data quality and standards, use of public data repositories, intellectual property rights, privacy of research participants, and long-term preservation and sustainability.

3.2 Guidance from Data Centres

3.2.1 UK Data Archive

The UK Data Archive (UKDA) provides archival services for the ESRC's Economic and Social Data Service, the National Centre for e-Social Science, the History Data Service and Census.ac.uk.[4] It has produced a series of Web pages to aid researchers in managing and sharing their data [Ukdb]. The following is a brief summary of the topics covered.

Confidentiality. The legislation relevant to confidentiality in the UK includes the Data Protection Act 1998, the Freedom of Information Act 2000, the Human Rights Act 1998, the Statistics and Registration Services Act 2007 and the Environmental Information Regulations 2004; there also exists in case law the notion of a duty of confidentiality. It is important that where data are likely to give rise to confidentiality issues, informed consent is sought not only for participation in research, but also the subsequent publication and sharing of the data. Various techniques exist for protecting confidentiality, such as anonymizing and aggregating data, applying access restrictions, and implementing secure data storage.

Copyright. Research works such as spreadsheets, reports and computer programmes fall under the definition of literary works and as such are subject to copyright. Ownership of copyright normally rests with the researchers and study participants, although the copyright for work completed under contract usually rests with the employer. Apart from copyright, there are also database rights, moral rights and publication rights to consider.

Metadata. The UKDA distinguishes three kinds of metadata. Study-level documentation includes data collection methods, lists of data sources, validation methods, and relationships to other datasets. Data-level documentation includes variables lists, coding and classification schemes, weighting and grossing methods, and explanations for missing data. Catalogue records provide metadata aimed at resource discovery and bibliographic

4. UK Data Archive Web site, URL: `http://www.data-archive.ac.uk/`

referencing; standards include Dublin Core, the General International Standard Archival Description, the Metadata Encoding and Transmission Standard and ISO 19115.

Formats and software. The UKDA specifies preferred formats for deposited data. It also provides advice on quality control, authenticity, version control, and audio file transcription. The UKDA uses a four-level classification system (A*, A, B, C) for judging the extent to which data have been cleaned and processed ready for sharing; the criteria for the A* level may serve as a checklist for researchers preparing data for deposit.

3.2.2 Archaeology Data Service

The Archaeology Data Service (ADS) was formerly part of the Arts and Humanities Data Service, but since the latter's closure it has been funded in its own right by the AHRC. It has produced an online guide for data depositors, covering the stages of starting a project, creating and documenting files, and documenting the project [Niv08]. A brief summary of the topics covered follows.

Starting the project. When planning a project, researchers should establish what types of data will be produced, and consider how these data may be re-used. The ADS provides guides on specific data types such as digital aerial photography interpretations, excavation and fieldwork data, Geographic Information Systems (GIS) and geophysics data. As a result of these considerations, a Digital Archive Strategy should be produced. A file naming convention should be established early on; the ADS recommends using only alphanumeric characters, the hyphen and underscore in filenames, reserving the full stop for demarcating the (lower case) file extension. Filenames should be unique within a collection.

Creating and documenting files. The ADS specifies preferred formats for deposited data, and in each case specifies what accompanying documentation is also required. For example, Scalable Vector Graphics (SVG) files should be accompanied by a caption and description of their relationships to other documents, while AutoCAD documents should be accompanied by an explanation of any conventions used (with respect to layers, colours, line types, hatching styles, etc.) and well as any relationships to databases, object libraries or other files.

Documenting the project. The ADS provides a template for providing metadata about an entire dataset. The template is a profile of Dublin Core metadata.

3.2.3 NERC data centres

NERC funds six designated data centres, each specializing in a different form of environmental data. These are the British Atmospheric Data Centre, the British Oceanographic Data Centre, the Environmental Information Data Centre, the National Geoscience Data Centre, the NERC Earth Observation Data Centre, and the Polar Data Centre. All of these offer bespoke guidance to projects on how to prepare data for deposit, but some also offer additional guidance between this level and the general guidance provided by NERC.

British Atmospheric Data Centre (BADC). The BADC accepts metadata principally packaged together with data in either NASA Ames Format for Data Exchange [GH98] or the

Network Common Data Form binary format, [Rew+10]. For tabular data, it recommends providing information about the experiment (instrument, spatiotemporal coverage), the experimenters, the independent variables, the files making up the set, the data format, the data version and processing level, and so on; other recommendations are provided for satellite imagery and software. It imposes a file naming convention consisting of the instrument name, the observation location, the date (and time) of the measurements, and additional information (e.g. version number, range resolution) all separated by underscores, and followed by a dot and extension [Bad].

British Oceanographic Data Centre (BODC). As well as bespoke advice, the BODC offers both a general guide for data submission and specific advice for particular data types [Bod]. Data may be submitted in most formats, so long as the format is either well supported or well documented; the way in which the format has been used should also be described. Parameters should be described in detail (column headings, units, whether measured directly or derived, how derived). The metadata should include where and when the data were collected, the collection method used, how the data are organized, how they have been processed, and points that subsequent users of the data should note.

NERC Earth Observation Data Centre (NEODC). The NEODC prefers to accept metadata in CSDGM format [Met98], as it currently uses this internally. For tabular data, it makes the same metadata recommendations as the BADC; other recommendations are provided for satellite imagery [Neo].

3.3 Digital Curation Centre

The Digital Curation Centre (DCC) has been tasked with providing guidance for projects on drawing up data management plans. To this end, Jones [Jon09a] provides an overview of funders requirements for sharing and preserving project outputs, both in terms of publications and research data, and examines the state of practice with regards curation policy provision. Jones provides the following general guidance on creating curation policies.

- Policies should fit around existing workflows and structures, so that the implementation of the policies leads to minimal changes in these workflows.
- If additional infrastructure is needed to comply with the policy, it should be put in place at the same time.
- The purpose of the policy should be clear, with the rationale for and expected benefits of the policy laid out explicitly. The policy should have a clear scope with regard to the data records covered, the activities being addressed, and the organizational unit to which it applies.

In addition, Jones [Jon10] provides a brief summary of the requirements of UK funders specifically relating to data management and sharing plans. Donnelly and Jones [DJ09] have combined these requirements along with additional input from practitioners and domain experts to produce a checklist of items to include within a data management plan. The DCC has also compiled a list of tools and resources that may be of use when drawing up data management plans and determining a suitable data repository with which to consult [Dcc].

3.4 Summary of guidance

The guidance given by funders, data centres and the digital curation community is broadly consistent. Independent of any particular research activity, there are some infrastructural questions to consider. Should the research centre adopt a global directory and file naming convention, or should there be several context-dependent conventions, or a different convention for each investigation? Is the management of institutional data storage sufficient to prevent corruption or loss of data, or should additional measures be taken by the research centre (e.g. a registry of checksums for determining file fixity)? How can the provenance of data be tracked (e.g. through version control systems, use of file spaces with limited permissions)? How will datasets be made available (if at all) and how will they be made citable? Will raw data be kept or just the finished, worked-up data?

The early planning stages for an investigation should consider the data that will be required to answer the research question. Do these data already exist, or will they need to be generated? In the case of new data, what sort of quantities are envisioned? Can these data be made consistent with existing data through the use of standard collection methods, standard data formats and standard metadata/packaging formats? What restrictions might need to be applied to the data, and are there any processes (e.g. anonymizing) that might enable sensitive data to be published openly? To what other purposes might this data be put? While it is not possible to predict all possible ways in which data might be re-used, it may be possible to work up data in such a way that it supports some obvious secondary uses as well the primary research. If software is to be produced, what development methodology will be used?

Particularly important at this early stage is to estimate the costs associated with working up the data to a reusable state and preserving it over the long term. On the latter point it is worth noting that due to the cost of storing a fixed amount of data halving every 14 months [Kom09], the total lifetime cost of storing data is unlikely to exceed twice the cost of storing it for its first 14 months. Once these costs have been determined, they should be taken account of when the grant application is made.

When using data in the course of research it is important to keep track of how they flow. Pre-existing data should be cited and the responsible parties acknowledged. Clarity should be sought over the intellectual property rights associated with newly generated data. The instruments, software and processes used to generate, record, derive, refine, aggregate and collate data should be documented, as should the relationships between different sets of data; common or otherwise easily reproducible processing steps/methods should be preferred.

Working up the data for final publication should be done towards the end of the project, preferably before the end of the project funding period or within a fixed time period afterwards. The data should be reviewed for quality prior to being made available or accepted into the archive, and a date should be set for revisiting the data to check for continued relevance and accessibility.

4 STANDARDS AND TOOLS FOR DATA CURATION

Generalist digital curation is an active research area, and as a result a wealth of tools, technologies and standards have been produced to assist those responsible for digital curation in an institutional setting.

4.1 Assessing data holdings

4.1.1 Data Audit Framework

Inspired by the DRAMBORA methodology for repository assessment (see section 4.3.5), the Data Audit Framework (DAF) was developed as a means for higher education institutions to recognize their data assets and, as a result, reconsider their policies, procedures and practices for managing them.[5] Being a framework, the DAF does not provide a strict method for performing a data audit, but rather a methodology that can be adapted for different institutional contexts [Jon09b; Jon+09]. The DAF involves four stages:

1. *Planning the survey.* One of the early tasks in the process is to arrange interviews with researchers. Timing the survey is crucial: in busy periods, researchers have less time for giving interviews, but conversely in quiet periods, researchers are more likely to go on leave. This stage is also the point at which the method and scope for the survey should be detailed. These will be heavily influenced by the aim of the survey; if the aim is to determine storage needs, the survey would naturally concentrate on producing a comprehensive (though not broadly detailed) asset inventory, whereas if the aim is to improve workflows, the survey would concentrate on drawing challenges, issues and good practice from data interviews. These priorities will affect the order of the following two stages, and whether they can run concurrently.
2. *Identifying and classifying assets.* This data collection stage focuses on creating an inventory of existing data assets. The purpose of this inventory, in most cases, is to record a representative sample of data assets in order the gauge the types and volume of data being held. For consistency, it is better to enter this stage with a clear idea of what constitutes a data asset for the purpose of the survey; most surveys conducted to date have included numerical data, statistics, output from experimental equipment, survey results, interview transcripts, databases, images and audiovisual files [Jon09b, p. 7]. It is also helpful to have in mind the granularity at which the assets will be recorded. Classifying data, in this context, refers to prioritizing data assets for more detailed consideration; the DAF methodology suggests classifying data assets according to whether they are vital, important or of minor importance to the institution, though other metrics may be used. The classification stage may not be necessary if sampling has already resulted in a manageable inventory, or if little further work on the inventory is planned. Data may be collected using desk research, questionnaires, wikis, or interviews.
3. *Assessing the management of data assets.* This data collection stage has two purposes. One is to gather a more comprehensive set of metadata for high priority

5. Data Audit Framework Web site, URL: `http://www.data-audit.eu/`

data assets to aid in their curation. The other is to assess how well these assets are being managed, and to determine what additional resources might be needed to maintain the value of the assets, or to increase the efficiency of their curation. The information for this stage is best collected using interviews, possibly using the data lifecycle as a framing device (see section 2.3).

4. *Reporting and recommendations.* The findings of the survey should be synthesized and presented in a report. As well as indicating what has been learned about the character of the data asset holdings discovered, the report should detail the challenges, issues and deficiencies in current practice identified by the survey and recommend improvements that could be made.

Sample forms are provided for paper-based audits, and an online tool is provided for completing the audit electronically. The methodology will in future be known as the Data Asset Framework.

4.1.2 Data Seal of Approval

The Data Seal of Approval (DSA) is an assessment developed by Data Archiving and Networked Services (DANS), the Dutch data archive.[6] The DSA may only be awarded following external assessment, but the quality guidelines used by the Assessment Board are freely available [SHH10]. The guidelines recognize that responsibility for archival-quality data is shared amongst three groups: producers for the quality of the research data themselves, the repository for the quality of data storage and availability, and consumers for the quality of data use. Underlying the guidelines are five criteria for the sustainable archiving of research data:

1. The research data can be found on the Internet.
2. The research data are accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data.
3. The research data are available in a usable format.
4. The research data are reliable.
5. The research data can be referred to.

There are sixteen guidelines in all, covering the deposition, archiving and subsequent use of research data:

1. The *data producer* deposits the research data in a data repository with sufficient information for others to assess the scientific and scholarly quality of the research data and compliance with disciplinary and ethical norms.
2. The *data producer* provides the research data in formats recommended by the data repository.

6. Data Seal of Approval Web site, URL: `http://www.datasealofapproval.org/`

3. The *data producer* provides the research data together with the metadata requested by the data repository.
4. The *data repository* has an explicit mission in the area of digital archiving and promulgates it.
5. The *data repository* uses due diligence to ensure compliance with legal regulations and contracts.
6. The *data repository* applies documented processes and procedures for managing data storage.
7. The *data repository* has a plan for long-term preservation of its digital assets.
8. Archiving takes place according to explicit workflows across the data life cycle.
9. The *data repository* assumes responsibility from the data producers for access to and availability of the digital objects.
10. The *data repository* enables the users to utilize the research data and refer to them.
11. The *data repository* ensures the integrity of the digital objects and the metadata.
12. The *data repository* ensures the authenticity of the digital objects and the metadata.
13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.
14. The *data consumer* must comply with access regulations set by the data repository.
15. The *data consumer* conforms to and agrees with any codes of conduct that are generally accepted in higher education and research for the exchange and proper use of knowledge and information.
16. The *data consumer* respects the applicable licenses of the data repository regarding the use of the research data.

4.2 Preservation metadata

4.2.1 PREMIS

The PREMIS Data Dictionary grew out of work by OCLC and RLG to flesh out the preservation metadata required by the OAIS Reference Model with elements from existing schemata [OCL02; Pre05; PRE08]. It is uses a data model with five entities: Intellectual Entities, Objects, Events, Agents and Rights. Of these, most of the metadata elements ('semantic units') are associated with the Object entity; none are associated with Intellectual Entities. While PREMIS does not prescribe any particular implementation, XML Schemata are provided for exchanging PREMIS metadata.[7]

7. PREMIS Web site, URL: `http://www.loc.gov/standards/premis/`

Several papers have been written on applying PREMIS in a repository context. Hitchcock et al. [Hit+07] map PREMIS metadata elements onto their principal sources, out of the person submitting the content, the repository software, a file format identification/ characterization tool, repository policy documents, the preservation service provider, and environment registries. Woodyard-Robinson [WR07] describes how PREMIS has been implemented across sixteen different repositories and projects, including the Koninklijke Bibliotheek, the (UK) National Archives, the National Digital Newspaper Program at the Library of Congress, and the National Digital Heritage Archive at the National Library of New Zealand. Dappert and Enders [DE08] describe how the British Library used PREMIS alongside METS (a packaging standard) and MODS (a bibliographic metadata standard) as part of the ingest process for electronic journals.

4.2.2 CAIRO

The CAIRO Project (Complex Archive Ingest for Repository Objects) was a two-year collaboration between the Bodleian Library (University of Oxford), the John Rylands Library (University of Manchester) and the Wellcome Library, ending in August 2008.[8] The aim of the Project was to simplify the task of adding collections of diverse born-digital content to a preservation repository, by integrating a suite of relevant digital curation tools into a simple, unified user interface [Tho08]. As part of the development work, the Project developed an extensible framework of content models for frequently encountered object types, expressed in terms of appropriate descriptive, preservation, structural and administrative metadata to be included in METS files. These content models defined the desired output from the proof-of-concept tool developed by the Project, although not all of them were supported by the conclusion of the Project. The tool was implemented on top of the Eclipse platform,[9] and used a modular architecture to allow format chracterization tools such as DROID and JHOVE to be plugged in as needed.[10] The tool continues to be developed as part of the futureArch Project.[11]

4.2.3 InSPECT

It was explained in section 2.2 that within the OAIS Reference Model, Representation Information is the information that someone needs in order to understand a Data Object. It is a highly useful concept to use when considering the eventual use of a digital object by a consumer; however, when a repository is using Representation Information in its own preservation actions, for example when migrating a file from one format to another, it typically has to make a selection from the available Representation Information, and the choice made has an effect on the information that remains available to the consumer. In such circumstances, it is perhaps more useful to consider the related concept of a digital object's *significant properties* – those aspects of the digital object itself which must be preserved over time in order for the digital object to remain accessible and meaningful. Developing and expounding this concept was the aim of InSPECT (Investigating the

8. CAIRO Project Web site, URL: `http://cairo.paradigm.ac.uk/`
9. Eclipse Web site, URL: `http://www.eclipse.org/`
10. CAIRO ingest tool development page, URL: `http://sourceforge.net/projects/cairo-ingest/`
11. FutureArch Project blog, URL: `http://futurearchives.blogspot.com/`

Significant Properties of Electronic Content over Time), an 18-month project concluding in March 2009 and led by the Centre for e-Research, King's College London.[12] Four case studies were conducted, in which the significant properties of raster images, emails, structured text and digital audio respectively were defined, and the extent to which these properties survived transformation from one representation format to another were measured. From these case studies, a generic framework for defining significant properties was produced [Kni08b], alongside a significant properties data dictionary and a document discussing the factors that influence decisions on which properties are significant [Kni08a; Kni08c]. While research data were not explicitly considered, the framework may prove of use when prioritizing the collection of Representation Information within the research workflow.

4.2.4 DDI

The Data Documentation Initiative (DDI) provides a standard for social science data combining both metadata requirements and packaging requirements. The Combined Life Cycle Model, used as a basis for version 3 of the DDI Specification, was introduced in section 2.3.3; as indicated above, this model suggested five principal sets of metadata: study conception, data collection process, logical structure of encoded data, physical structure of encoded data, and archiving.

The DDI Specification itself is large and complex, consisting of 22 DDI XML schemata, a Dublin Core XML schema and 18 XHTML schemata. The DDI XML schemata may be broken down into four categories.

The five *Scheme-Based Modules* correspond to the five principal sets of metadata mentioned above. The *Conceptual Component* module deals with the subject matter of the data: what was measured, how it was grouped and which concepts were being tested. The *Data Collection* module corresponds to the process by which the data were collected, and hence deals with research methodology, survey instruments and data processing. The *Logical Product* module considers the data as a collection of facts, and deals with the parameters measured or controlled, the categories used, and the dependencies between the data. The *Physical Data Product* module deals with how the logical product is represented as a file: the file format, how records are structured within a database, etc. Finally, the *Archive* module deals with the organizations storing and curating the data, giving details on how the data are managed and how they may be accessed.

These modules are known as 'scheme-based' because as well as per-instance metadata they also define structures ('schemes') for describing entities that may apply to many datasets. For example, the Archive module defines the Organization scheme, the Data Collection module defines the Control Construct, Interviewer Instruction and Question schemes, and the Logical Product module defines the Category, Code, Variable and NCube (n-dimensional table) schemes.

There are also three *Non-Scheme Modules* which provide additional metadata. The *Comparative* module enables aspects of different studies to be compared while *DDI*

12. InSPECT Project page at KCL, URL: `http://www.kcl.ac.uk/iss/cerch/projects/completed/inspect`. InSPECT Project Web site, URL: `http://www.significantproperties.org.uk/`

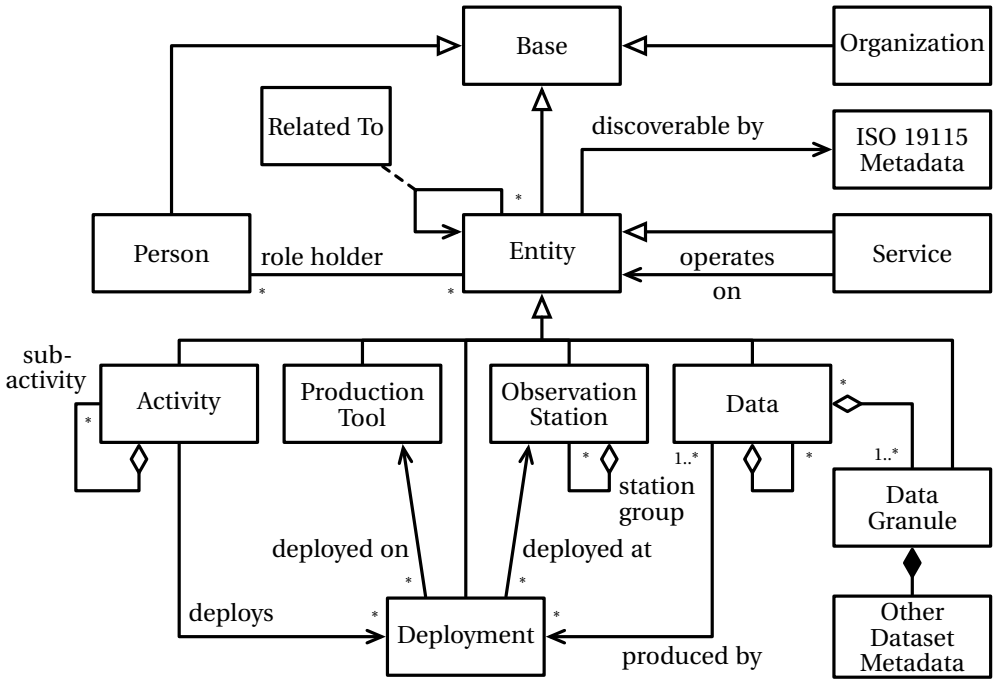


Figure 7: High level data model from the MOLES metadata profile, adapted from Lawrence et al. [Law+09].

Profile describes any additional conventions used in the metadata. The *Physical Instance* module describes a particular instance (file) of the data product: how it was created, how it was checked, fixity checksums, etc.

The five *Sub-Modules* pre-define common types of record layout, and specialist metadata for describing data in those layouts. There are also six types of *Shared Content*: *Reusable* contains common elements that would otherwise occur in several modules, while interface schemata are provided for compatibility with Dublin Core, XML and XHTML.

The remaining three schemata are the *Packaging Modules*. All data relating to a single study would be packaged into a *Study Unit*. If there are close similarities between several Study Units, they may be aggregated within a *Group* module. Finally, an entire programme of research, consisting of several Study Units or Groups thereof, may be packaged within a *DDI Instance*.

4.2.5 MOLES

The MOLES metadata profile, which underlies the NERC Data Grid, uses a data model with five key entities (see Figure 7). The *Deployment* entity represents a data gathering exercise, and links four other entities. The *Activity* entity represents the study or project under whose auspices the data gathering exercise is performed. The *Observation Station* entity represents the location(s) where the data is gathered, and is also used to record the researchers generating the data. The *Production Tool* represents the instrument and methodology used to collect the data. Finally, the data itself is represented by a *Data* entity; this entity is made up either of further Data entities, or of one or more *Data*

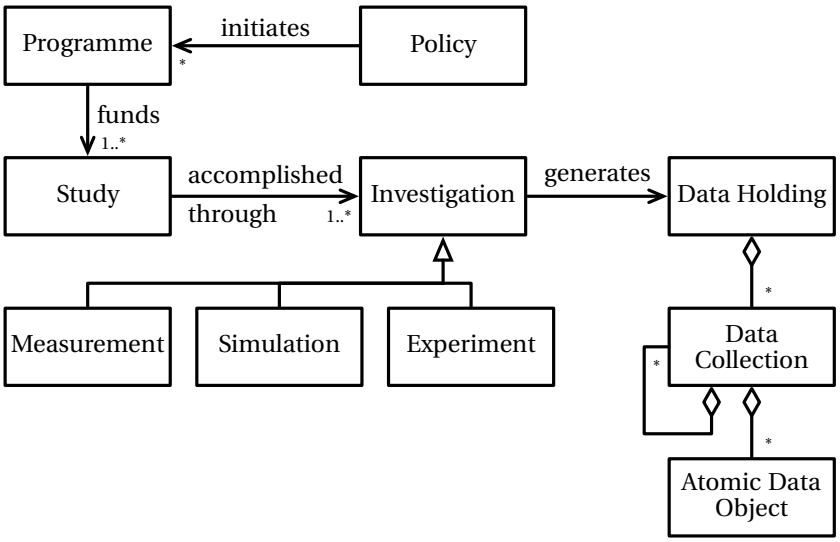


Figure 8: High level entity model from the CCLRC Scientific Metadata Model.

Granule entities, themselves essentially atomic datasets [Law+09]. MOLES allows for the metadata for Data Granule entities to be supplemented with more specific (archiving-related) metadata in a different scheme, such as CSML [Woo+06].

One further entity, *Service*, represents a process which can manipulate the other high-level entities to produce either new high level entities or text/visualizations. All high-level entities may be described using metadata from another scheme such as DIF [GCM08] or ISO 19115 [ISO03b], and their inter-relationships clarified by *Related To* associations.

4.2.6 CCLRC Scientific Metadata Model

The Scientific Metadata Model (SMDM) devised by STFC (formerly CCLRC) has several different models associated with it. The scientific activity model has four levels of entity. *Policy* is a governmental or company policy that drives research by initiating one or more *Programmes* of work. Each Programme represents a tranche of funding for studies or projects on a particular theme or topic. A *Study* is a piece of work performed by a principal investigator and/or institution, along with co-investigators and researchers. A Study is typically funded by a Programme, and therefore may have a grant number associated with it. An *Investigation* is a data collection exercise performed as part of a Study. The model explicitly recognizes three types of Investigation, although it leaves room for others.

- An *Experiment* typically consists of a controlled environment, where an instrument is used to measure one property of the environment or specimen while other properties are set to a known value or a series of known values.
- A *Measurement* is typically produced by a passive detector that records the state of the environment at specified intervals of time and space.

- A *Simulation* takes a mathematical model of a system, and from a set of initial parameters either calculates what a further set of parameters must be, or determines how the modelled system evolves over time.

The model also recognizes the existence of *Virtual Studies*. These are groups of studies that are related in some way – typically having the same principal investigator/institution and subject matter – other than belonging to the same Programme. Common examples are studies where one is the follow-on of the the other.

The collected data itself is covered in a different model. Each Investigation produces exactly one *Data Holding*. This Data Holding is made up of one or more *Data Collections*, each of which may be divided into further Data Collections. The concept of a Data Collection enables different sequences of data to be separated out, e.g. the raw data instrument data from the intermediate and processed sets of data. Each Data Collection is ultimately represented by a set of *Atomic Data Objects:* physical data files or database queries from which the data may be obtained.

These two models are combined and illustrated in Figure 8.

4.2.7 Dublin Core Application Profiles

Application profiles are '[metadata] schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application' [HP00, par. 2]. They are a way of creating a new metadata schema that is explicitly compatible (at least partially) with existing schemata, while also being specifically tailored for a given application. Typically this will involve

- selecting a pertinent subset of elements from an existing schema;
- combining several existing schemata that are each at least partially relevant for the application;
- refining elements from existing schemata, either to provide a narrower meaning in the context of the application, or to control the vocabulary used;
- filling in any gaps with a custom schema, to be separately maintained.

The Dublin Core Metadata Initiative (DCMI) has developed this idea further as part of its effort to refactor Dublin Core Metadata in terms of Linked Data [BL09]. It has produced an abstract model for metadata [Pow+05], and a framework known as the Singapore Framework for using this model to develop a Dublin Core Application Profile (DCAP) [NBJ08]. The definition of a DCAP is rather more complex and prescriptive than that of application profiles generally (see Figure 9); the Framework defines a DCAP as a packet of three to five components:

- *functional requirements (mandatory):* the application that the profile is intended to support, in terms of specific functions that are in (or out) of scope;
- *domain model (mandatory):* the basic entities described by the profile, and the relationships between them.
- *description set profile (mandatory):* a set of rules that define what constitutes a valid instance of the application profile.

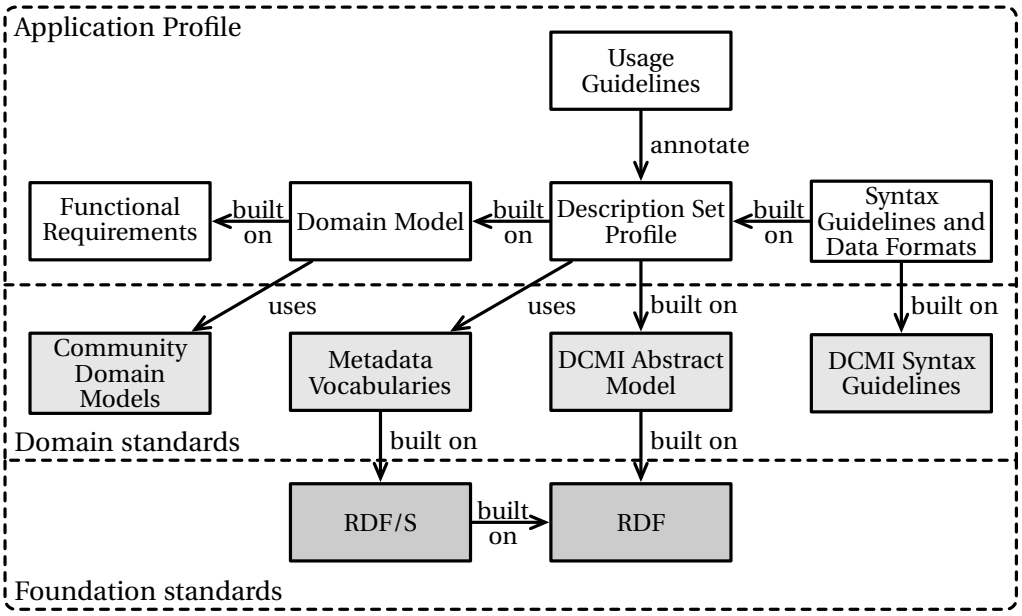


Figure 9: Singapore Framework for Dublin Core Application Profiles.

- *usage guidelines (optional):* human-orientated information and guidance on using the application profile in practice.
- *encoding syntax guidelines (optional):* any additional syntactical rules relating to the application that aren't covered by the description set profile and the normal rules of the language used to express instances of the application profile.

In addition, the Framework recommends that the domain model for a DCAP is based on an existing, widely used domain model, and that any syntax guidelines should take into account those provided by DCMI. The Framework requires that the description set profile should specify how entities of the DCMI Abstract Model are used in the DCAP, and by extension how the DCAP may be expressed in RDF. Any metadata vocabularies used by the description set profile should be expressed using RDF Schema, otherwise known as RDF Vocabulary Description Language.

One of the first DCAPs to be designed for scientific data was the Dryad Application Profile, developed by the University of North Carolina Metadata Research Center and the (US) National Evolutionary Synthesis Center for use in a repository of evolution and ecology research [Gre+09]. Version 2 of the profile defines two modules for describing citations and data objects respectively, with elements drawn from Dublin Core, Darwin Core, Publishing Requirements for Industry Standard Metadata (PRISM), the Journal Publishing Tag Set Tag Library, DDI, Ecological Metadata Language (EML) and PREMIS.

In the UK, the JISC commissioned a scoping study to determine the feasibility and utility of producing an application profile for the discovery and delivery of scientific datasets [Bal09]. The study examined 15 metadata schemes and profiles already in use for describing scientific data, and identified 33 metadata elements each of which were used in at least 3 of the schemes, and 16 of which were used in at least 10 schemes. The report concluded that there is scope for an application profile dedicated to the discovery

and delivery of data sets, but warned that more specialist metadata would be required for discipline-specific applications.

4.2.8 Software preservation

Software Preservation: Standards is a research project of the STFC e-Science Centre.[13] It investigates the costs and benefits for software repositories of using standards such as the OAIS Reference Model and PREMIS, while monitoring and helping to develop new and existing standards. The project has been informed by a case study looking at the Verified Software Repository [BHW06], and how the code for the Repository is itself managed on the SourceForge software repository.[14]

4.3 Archive management tools

4.3.1 AIDA

AIDA (Assessing Institutional Digital Assets) is a self-assessment tool that institutions can use to assess their current capabilities and capacity with regard to digital preservation. It uses a balanced scorecard approach across three different dimensions:

- Viewpoint: Organisational, Technology, Resources.
- Stage of development: Acknowledge, Act, Consolidate, Institutionalize, Externalize.
- Administrative level: Institution, Department.

The toolkit identifies elements within each viewpoint that are important for digital preservation, then provides matrices for judging at which stages of development the institution and department are, respectively, in the context of each element. Each matrix contains descriptions of what each stage means in the context of the element and administrative level, along with indicators and exemplars.

Currently, completed scorecards must be returned to the AIDA Project for interpretation, although some insight can be gained merely from performing the exercise. In due course, the toolkit will suggest strategies for institutions and departments based on their level of readiness, recommend improvements, provide tailored tools for assessing data assets, and estimate preservation costs.

4.3.2 DCC Methodology for Designing and Evaluating Curation and Preservation Experiments

Having decided on a curation strategy, it is important to know that it is fit for purpose. When dealing with long-term preservation, there is a limit to the extent to which a strategy can be tested given that the technical capabilities and working knowledge of the

13. *Software Preservation: Standards* Web site, URL: `http://www.e-science.stfc.ac.uk/projects/software-preservation/software-preservation-standard.html`
14. Verified Software Repository development page, URL: `http://vsr.sourceforge.net/`

future Designated Community are unknown, but there are some tests one can perform to ensure basic operations are succeeding. For example, if the strategy involves normalizing files to a particular format, one can check that the normalization process produces valid, well-formed files containing all the significant information. One could also check to see if a dataset is sufficiently documented by asking a member of the current Designated Community if they can understand it.

The DCC has produced a methodology for conducting such experiments, based largely on a methodology devised by the Planets project and incorporating elements of others produced by the CASPAR project, the DELOS project and a team led by the (Dutch) Nationaal Archief, respectively [JMK09; Kim08]. It follows an eight-stage process:

1. Develop use case scenario. This is a use of the data that the preservation action is supposed to support and preserve. The use case should specify a digital resource, why and how it is used, by whom, within which Designated Community, and with what frequency.
2. Define basic properties. Such properties typically include the name, description, purpose and scope of the experiment.
3. Design experiment. The details of the experimental method – choice of sample, processing stages, etc. – are determined at this point.
4. Specify outcomes. Specifically, the criteria for judging the success of the experiment should be specified. If it is not obvious how the curated resource's ability to satisfy the use case in stage 1 could be judged, this should be specified alongside any technical outcomes.
5. Go or no go decision. At this stage the experiment may be halted, postponed or allowed to proceed, depending on feasibility.
6. Run experiment.
7. Evaluate results. This includes a technical assessment of the curation action, judged using the criteria specified in stage 4 and typically involving a comparison of the input and curated resources. It should also include a qualitative assessment based on implementing the use case with the curated resource.
8. Publish the results in a DCC report. The wording of this stage reflects the genesis of the methodology, but the idea is to report back findings to the preservation community in order to build up a shared evidence base concerning preservation and curation techniques.

4.3.3 OpenDOAR Policies Tool

The OpenDOAR Policies Tools was developed in response to research indicating that two thirds of Open Access repositories did not have policies for content submission, re-use, preservation, etc. [Mil06]. It provides a set of templates for metadata, data, content, submission and preservation policies, and a simple interface for choosing between the given alternatives.[15] Policies created using the tool can be output as plain text, HTML, or as EPrints source code for static pages or configuration files. The latter files allow the policies to be accessed through the OAI-PMH protocol.

15. OpenDOAR Policies Tool Web page, URL: `http://www.opendoar.org/tools/en/policies.php`

4.3.4 DINI-Zertifikat

The Deutsche Initiative für Netzwerkinformation (DINI) Certificate Document and Publication Services [DIN06] provides a standard of quality for higher education institutional repositories. It provides both minimum standards and recommendations for the visibility of services; repository policies; support provided for authors; handling of copyright and licensing issues; security, authenticity and data integrity of both the repository and individual documents; subject indexing, metadata export and repository interfaces; logs and statistics; and long-term availability. The certificate itself is awarded after external inspection.

4.3.5 DRAMBORA

DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) is a self-assessment toolkit developed by the Digital Curation Centre and DigitalPreservationEurope.[16] The toolkit may be used interactively online or downloaded for use in paper form. The audit has six stages, in which the auditor identifies the organization's role and objectives, policy framework, activities and assets, before identifying and assessing the risks associated with these activities and assets, and developing a strategy to manage them [IV09].

4.3.6 TRAC

Trusted Repositories Audit and Certification (TRAC) has its origins in a 2002 report on the attributes and responsibilities of trusted digital repositories, written by a joint working group of OCLC and RLG [RLG02]. This latter report fulfilled its brief at a rather abstract level, and recommended further work to develop a certification programme using detailed and specific criteria. As a result, RLG and the (US) National Archives and Records Administration (NARA) set up an international task force for digital repository certification, producing a draft audit checklist in 2005 [RLG05] and a final version of the TRAC Criteria and Checklist in 2007 [RLG07]. The criteria are divided into three sections, relating to organizational infrastructure (governance, sustainability, staffing, etc.), digital object management (OAIS functions), and technical infrastructure (system architecture, technologies, security) respectively. TRAC is now maintained by the Center for Research Libraries (CRL).

4.3.7 PLATTER

The Planning Tool for Trusted Electronic Repositories (PLATTER) is a strategic-level tool devised by DigitalPreservationEurope for ensuring trustworthiness emerges as a characteristic of a digital repository [Ros+08]. The tool begins with a questionnaire for determining the character of the repository in question, then provides a framework for defining goals and performance targets for the repository, and advice on expanding these

16. DRAMBORA Web site, URL: `http://www.repositoryaudit.eu/`

into a set of nine Strategic Objective Plans. These plans cover, respectively, financial monitoring, staffing, data/metadata, acquisition of content, access to content, preservation of content, technical systems, succession and disaster planning. The aim for PLATTER is that a repository developing and implementing its own version of the Strategic Objective Plans should find itself performing well in trustworthiness audits.

4.4 Costs and benefits

For curation, and preservation in particular, to be effectively undertaken requires a not inconsiderable outlay in terms of both staff time and equipment. It is also a long-term commitment on the part of an institution. If curation activities are to receive the sustained investment needed, institutions need to be convinced first of all that there is either a net financial benefit to performing curation or a worthwhile return for the additional cost, and second of all that the chosen architecture for curation represents the best value for money for the institution. Making either case requires a framework for assessing the costs and benefits of curation. The four main frameworks developed so far for this purpose are *espida*, LIFE, Keeping Research Data Safe and the *Identifying Benefits* report.

4.4.1 ESPIDA

The *espida* (Effective Strategic model for the Preservation and disposal of Institutional Digital Assets) Project ran from October 2004 to January 2007, and in that time developed an approach to the process of funding projects.[17] In the higher and further education context in particular, there is a tendency for projects to produce new knowledge, new information or improved processes rather than financial rewards. This makes it hard for proposers to present a project's risks and rewards in anything other than vague terms, and makes it hard for the funders to set these risks and rewards against the requested investment. The *espida* Approach is an attempt to provide a more objective and concrete way of expressing the intangible outcomes of projects. It encourages proposers to align their proposals to the funder's strategic goals, and uses Outcome Scorecards and cost templates to express benefits and costs. The *espida* Handbook includes a case study using the Approach in an institutional repository setting [CM07, pp. 36–42].

4.4.2 LIFE

The LIFE (Lifecycle Information For E-literature) Project is a collaboration between University College London (UCL) and the British Library.[18] It is looking at the lifecycle of digital material, with particular regard to the cost of collecting and preserving the material.

The first phase of the Project ran for one year ending in April 2006; it used UCL and British Library collections to develop a model of the lifecycle of a digital object, and a methodology for costing each stage in that lifecycle [MWA06]. Within the second phase,

17. *espida* Project Web site, URL: `http://www.gla.ac.uk/espida/`
18. LIFE Project Web site, URL: `http://www.life.ac.uk/`

which ran for 18 months and ended in August 2008, the lifecycle model was refined and three further costing case studies were produced: SHERPA-LEAP, SHERPA-DP and the Burney Collection [Ayr+08]. The third phase began in August 2009 and will last one year. The focus of this phase will be the further refinement of the Generic Preservation Model and to develop a predictive costing tool, available both as a spreadsheet and as a Web application.

4.4.3 Keeping Research Data Safe

Beagrie, Chruszcz and Lavoie [BCL08] propose a framework for determining the medium to long term costs to higher education institutions (HEIs) of preserving research data.[19] The framework uses full economic costing, in order to support more accurate cost–benefit analysis and more accurate comparisons between in-house and outsourced solutions, and is tailored towards use with the Transparent Approach to Costing (TRAC) in use in UK HEIs. It consists of three parts: a list of key cost variables and units that affect how preservation costs change over time; an activity model identifying the activities required to preserve research data, each of which having a cost implication; and a resources template providing TRAC-orientated cost category headings into which the costs identified by the activity model should be separated. The framework was developed with reference to the LIFE cost models, the OAIS Reference Model and the NASA Cost Estimation Tool [Fon+07], alongside four case studies of working data archives.

The framework proposed by Beagrie, Chruszcz and Lavoie was further refined in the Keeping Research Data Safe 2 Project, which ran for ten months from March 2009. A wider data survey was conducted to corroborate the findings of the earlier report, and the activity model was reviewed and modified accordingly.[20]

4.4.4 Identifying Benefits

Fry et al. [Fry+08] identify the benefits arising from curating and openly sharing research data.[21] Among the direct benefits identified were: the potential for new discoveries from existing data, reduced duplication of data collection costs, increased transparency of the scientific record, and higher and more rapid impact of scientific research. Fry et al. also identified potential for pedagogical uses of research data as an indirect benefit.

Their report contains a methodology for performing a cost–benefit analysis of curating research data. On the costs side it extends the KRDS model [BCL08] to take account of costs incurred by those depositing and re-using data. On the benefits side, it focuses on calculating direct and indirect cost savings, while providing some scope for calculating increased returns and other more distant benefits. Appendix 3 of the report provides a worked example of a cost–benefit analysis for an institutional data repository.

19. *Keeping Research Data Safe* project page on the JISC site, URL: `http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx`
20. Keeping Research Data Safe 2 Project Web page, URL: `http://www.beagrie.com/jisc.php`
21. *Identifying Benefits* project page on the JISC site, URL: `http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/databenefits.aspx`

5 CURATION IN PRACTICE

5.1 World Data Center for Climate

In 2007–2008, the International Council for Science's World Data Center for Climate (WDCC) implemented a new data management strategy, in response to rising total storage costs [Lau08; LS07]. Instead of running a single long-term archive for all project data, a four-tier system was introduced. Projects working with the data centre are allocated three types of disc space. The first is temporary disc space on which calculations and other forms of data processing may be done; this space is not used for storage at all, as once the calculations are complete and the results retrieved, the files are deleted. The second is working storage, used by the project during its lifetime but not beyond. The third is archival storage, where data files of lasting value are transferred in order to be prepared for long-term storage; files stored here are deleted one year after the project ends. The fourth tier of storage is the long-term ('documented') data archive, where data from the archive tier are transferred once they have been sufficiently worked up and checked for quality. The long-term archive holds data for at least ten years.

The WDCC performs three levels of quality assurance tests. First, a semantic check is performed during the lifetime of the project to determine the validity and usefulness of the data: passing this is a pre-requisite for data to be transferred to the short-term archive. Secondly, data documentation and syntactic proof routines are conducted as part of the ingest process into the long-term archive. The third level relates to the creation of reference data within the wider dataset: these are the data that researchers would cite in publications. Once a set of reference data has been identified, it and its accompanying metadata are subjected to detailed review to ensure they are widely re-usable. On passing these checks, the set is assigned an Scientific and Technical Data Digital Object Identifier (STD-DOI) and a catalogue record produced; this record is used not only within the WDCC catalogue but also made available for harvest by other catalogue systems.

This case study contains several points that the ERIM Project may wish to consider when constructing and implementing its own data management plans.

1. Curating data properly is a non-trivial investment in both the short and medium term, so it should be performed with discrimination. It is best to be clear early on about the kinds of data that will be treated as ephemeral and the kinds that will be preserved in the long term.
2. Observing a time limit for working up data helps to ensure that it does get done.
3. Setting firm standards for data quality in advance aids both those working up the data and those checking it for quality.
4. Scheduled reviews for archived data provide opportunities (*a*) to dispose of obsolete data, and (*b*) where data remain relevant, to ensure that the accompanying metadata are still adequate for contemporary researchers.

5.2 University of Bath

Within the context of research performed during the requirements gathering stage of the ERIM Project, initial discussions with engineering researchers were used as the basis for

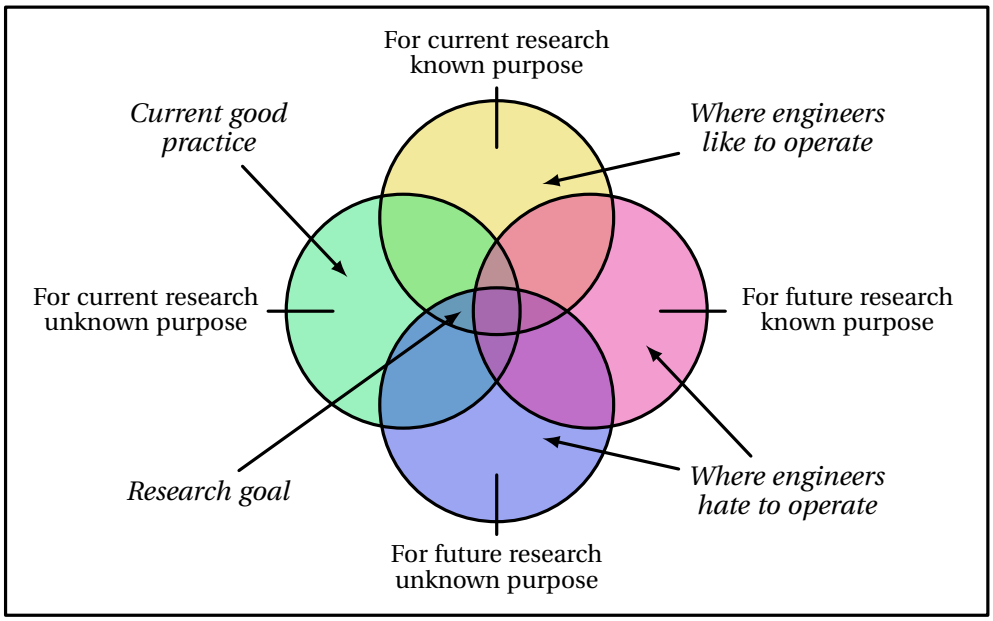


Figure 10: Venn diagram mapping different modes of data management.

a model describing the modes of engineering research data management with respect to re-use and re-purposing (see Figure 10).

The impression gained by these initial discussions was that the predominant mode of data management among researchers is in the gathering and organization of data for immediate use in the course of the present research, falling into the upper circle of the diagram. In cases of good management practice researchers will manage their data so that others within the same area of research can re-examine them reliably at a later date, perhaps to verify the earlier results or compare them with data gathered elsewhere; such management practice is represented by the left-hand circle. In engineering research, data are rarely managed in either of the modes represented by the lower and right-hand circles. The mode represented by the right-hand circle is a rather special case where it is known in advance how a different set of researchers will want to use the data, so the first set of researchers manage their data in the light of this knowledge. The mode represented by the lower circle is most often associated with data centres and repositories rather than researchers themselves, though way in which the researchers manage their data impacts strongly on the management that can be performed *post hoc*.

In terms of this model, the proposed research goal for ERIM can be seen as providing data management plans that work in the intersection of the modes represented by the upper, left-hand and lower circles. Here plans will help researchers manage their data in such a way that they will be more reliably and meaningfully reusable for future research and by other researchers.

6 CONCLUSIONS AND RECOMMENDATIONS

One of the principles underlying the curation of research data is that data are most useful when they are interoperable with other data, and that the best way to achieve this is by adherence to a common standard or set of standards. What is true for data is also true for data management: use of common tools and terminology enables greater understanding and co-operation between data managers and data curators at different centres. It is therefore a recommendation of this report that existing terminology and ready-made tools should be used wherever possible.

The terminology of the Open Archival Information System (OAIS) Reference Model is strongly recommended as the framework in which data management plans should be presented. The terminology of the DCC Curation Lifecycle Model is recommended as a high-level view of the entire data lifecycle, although it is acknowledged that ERIM's focus will be on the first few stages of that lifecycle, and therefore that a finer-grained approach will be needed. It is recommended that detailed modelling of data flows and relationships should be performed using an existing formal modelling technique such as IDEF or UML.

In the course of this report, points of significance for the ERIM Project were drawn from consideration of the OAIS Reference Model (section 2.2.1), guidance provided by funding bodies, data centres and the digital curation community (section 3.4) and from data curation in practice (section 5.1). These points should be borne in mind when drawing up data management strategies and plans.

The Data Audit Framework is recommended in the context of the ERIM case studies; given that an earlier audit has already fed into the selection of the case studies, any further auditing should omit the classification process and instead concentrate on assessing the management of data assets. The Data Seal of Approval is commended as a checklist of points to consider when drawing up data management strategies and plans.

On the matter of metadata supporting the curation of research data, PREMIS is recommended as a good set of preservation metadata. For descriptive and management metadata, ERIM is encouraged to consider the elements of the Dryad Application Profile and the elements identified by the Scientific Data Application Profile Scoping Study. Furthermore, when considering the metadata needed for re-use, the technique of considering significant properties is highly recommended.

When drawing up a data management plan, the OpenDOAR Policies Tool and the DCC Checklist for Data Management Plans are recommended as starting points. If time permits, it is recommended that ERIM performs at least one preservation experiment to ensure that the plan is fit for purpose.

REFERENCES

[AHR09] Arts and Humanities Research Council (2009-12). *Research Funding Guide*. Version 1.8. Arts and Humanities Research Council: Bristol. URL: `http://www.ahrc.ac.uk/FundingOpportunities/Documents/Research%20Funding%20Guide.pdf` (2010-06-01).

[Ayr+08] P Ayris et al. (2008-08-26). *The LIFE^2 Final Project Report*. Final Report. JISC. URL: http://eprints.ucl.ac.uk/11758/ (2009-10-13).

[Bad] *BADC Help Page* (2009-03-18). British Atmospheric Data Centre. URL: http://badc.nerc.ac.uk/help/ (2010-03-25).

[Bal09] A Ball (2009-06-03). *Scientific Data Application Profile Scoping Study*. Project report. Version 1.1. JISC. URL: http://www.ukoln.ac.uk/projects/sdapss/papers/ball2009sda-v11.pdf (2009-11-24).

[BBS07] Biotechnology and Biological Sciences Research Council (2007-04). *Data Sharing Policy*. Biotechnology and Biological Sciences Research Council: Swindon. URL: http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.pdf (2010-06-01).

[BCL08] N Beagrie, J Chruszcz & B Lavoie (2008-04). *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. Final report. HEFCE. URL: http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf (2009-10-13).

[Bea06] N Beagrie (2006-11). 'Digital Curation for Science, Digital Libraries, and Individuals'. *International Journal of Digital Curation* 1:1. ISSN: 1746-8256. URL: http://www.ijdc.net/ijdc/article/view/6 (2009-11-16).

[BHW06] J Bicarregui, CAR Hoare & J Woodcock (2006). 'The Verified Software Repository: A Step Towards the Verifying Compiler'. *Formal Aspects of Computing* 18:2, 143–151. ISSN: 0934-5043. DOI: 10.1007/s00165-005-0079-4. URL: http://epubs.cclrc.ac.uk/work-details?w=33971 (2010-07-01).

[BL09] T Berners-Lee (2009-06-18). *Linked Data*. World Wide Web Consortium. URL: http://www.w3.org/DesignIssues/LinkedData.html (2010-03-16).

[Bod] *Submitting data to BODC* (2010-03-24). British Oceanographic Data Centre. URL: https://www.bodc.ac.uk/data/data_submission/ (2010-03-25).

[BR06] R Bose & F Reitsma (2006-05). *Advancing Geospatial Data Curation*. Institute of Geography Online Paper GEO-013. School of Geosciences, University of Edinburgh. URL: http://hdl.handle.net/1842/1074 (2009-11-11).

[BT09] A Burton & A Treloar (2009). 'Designing for Discovery and Re-Use: the "ANDS Data Sharing Verbs" Approach to Service Decomposition'. *International Journal of Digital Curation* 4:3, 44–56. ISSN: 1746-8256. URL: http://www.ijdc.net/ijdc/article/view/133/ (2010-03-11).

[Bun+06] P Buneman et al. (2006-05). 'A Provenance Model for Manually Curated Data'. In: *Provenance and Annotation of Data*. Vol. 4145. Lecture Notes in Computer Science. Springer: Berlin & Heidelberg, 162–170. ISBN: 978-3-540-46302-3. DOI: 10.1007/11890850.

[CCS02] Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. Blue Book CCSDS 650.0-B-1. Also published as ISO 14721:2003. URL: http://public.ccsds.org/publications/archive/650x0b1.pdf.

[CM07] J Currall & P McKinney (2007-01-23). *espida Handbook: Expressing project costs and benefits in a systematic way for investment in information and IT*. University of Glasgow. URL: http://hdl.handle.net/1905/691 (2009-10-13).

[Com86] Committee on Data Management and Computation (1986). *Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences*. Space Science Board, National Research Council. National Academy Press: Washington, DC. URL: http://www.nap.edu/openbook.php?record_id=12343&page=32 (2010-03-26).

[DCC07] Digital Curation Centre (2007-04-26). *What is Digital Curation?* URL: http://www.dcc.ac.uk/about/what/ (2009-11-11).

[Dcc] *Policy Tools and Guidance* (2010). Digital Curation Centre. URL: http://www.dcc.ac.uk/resources/policy-and-legal/policy-tools-and-guidance (2010-03-25).

[DE08] A Dappert & M Enders (2008). 'Using METS, PREMIS and MODS for Archiving eJournals'. *D-Lib Magazine* 14:9/10. ISSN: 1082-9873. DOI: 10.1045/september2008-dappert.

[DGS09] D De Roure, C Goble & R Stevens (2009-05). 'The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows'. *Future Generation Computer Systems* 25:5, 561–567. ISSN: 0167-739X. DOI: 10.1016/j.future.2008.06.010.

[DIN06] DINI Electronic Publishing Working Group (2006-09). *DINI-Certificate Document and Publication Services 2007*. DINI Schriften 3-en. Version 2.0. Deutsche Initiative für Netzwerkinformation. URL: http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf (2010-02-17).

[DJ09] M Donnelly & S Jones (2009-06-17). *Data Management Plan Content Checklist: Draft Template for Consultation*. Digital Curation Centre. URL: http://www.dcc.ac.uk/sites/default/files/documents/templates/DMP_checklist.pdf (2010-03-25).

[Ear86] Earth Observing System Data Panel (1986). *Earth Observing System: Data and Information System*. Technical Memorandum. Version 87777. National Aeronautics & Space Administration. URL: http://hdl.handle.net/2060/19860021622 (2010-03-25).

[EPS06] Engineering and Physical Sciences Research Council (2006). *Guide to Good Practice in Science and Engineering Research*. Engineering and Physical Sciences Research Council: Swindon. URL: http://www.epsrc.ac.uk/CMSWeb/Downloads/Other/GoodPracticeGuideSciEngRes.pdf (2010-06-01).

[Esd] *Advice for Je-S applicants* (2009-09-14). Economic and Social Data Service. URL: http://www.esds.ac.uk/aandp/create/esrcfaq.asp (2010-03-24).

[ESR00] Economic and Social Research Council (2000-04). *Data Policy*. Economic and Social Research Council: Swindon. URL: http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Images/DataPolicy2000_tcm6-12051.pdf (2010-01-11).

[Fis09] P Fisher (2009-11-20). *Pathways and Gene Annotations for QTL Region*. Version 5. myExperiment. URL: http://www.myexperiment.org/workflows/16?version=5 (2010-03-12).

[Fon+07] K Fontaine et al. (2007-10). 'Observations on Cost Modeling and Performance Measurement of Long-Term Archives'. In: *PV 2007: Ensuring the Long Term Preservation and Value Adding to Scientific and Technical Data – Conference Proceedings*. Ed. by E Mikusch & C Reck. German Aerospace Centre (DLR) & German Remote Sensing Data Centre (DFD): Oberpfaffenhofen, Germany. ISBN: 978-3-00-022548-2. URL: http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf (2009-10-21).

[Fry+08] J Fry et al. (2008-11). *Identifying Benefits Arising from the Curation and Open Sharing of Research Data Produced by UK Higher Education and Research Institutes*. Final report. JISC. URL: http://ie-repository.jisc.ac.uk/279/ (2009-10-13).

[GCM08] Global Change Master Directory (2008). *Directory Interchange Format (DIF) Writer's Guide*. National Aeronautics & Space Administration. URL: http://gcmd.nasa.gov/User/difguide/difman.html (2008-11-26).

[GH98] SE Gaines & RS Hipskind (1998-06-18). *Format Specification for Data Exchange*. Version 1.3. National Aeronautics & Space Administration. URL: http://cloud1.arc.nasa.gov/solve/archiv/archive.tutorial.html (2010-03-25).

[Gia07] D Giaretta (2007-10). 'Preservation Concepts and the Role of This Conference Series'. In: *PV 2007: Ensuring the Long Term Preservation and Value Adding to Scientific and Technical Data – Conference Proceedings*. Ed. by CR Eberhard Mikusch. German Aerospace Centre (DLR) & German Remote Sensing Data Centre (DFD): Oberpfaffenhofen, Germany. ISBN: 978-3-00-022548-2. URL: http://www.pv2007.dlr.de/Papers/Giaretta_Keynote.pdf (2009-11-11).

[GK02] A Green & JP Kent (2002). 'The Metadata Life Cycle'. In: *MetaNet Work Package 1: Methodology and Tools*. Ed. by JP Kent. Chap. 2.2, 29–34. ISBN: 1-85764-017-9. URL: http://www.epros.ed.ac.uk/metanet/deliverables/D4/IST_1999_29093_D4.pdf (2010-03-11).

[Gre+09] J Greenberg et al. (2009). 'A Metadata Best Practice for a Scientific Data Repository'. *Journal of Library Metadata* 9:3/4, 194–212. ISSN: 1938-6389. DOI: 10.1080/19386380903405090.

[Hig08] S Higgins (2008-07). 'The DCC Curation Lifecycle Model'. *International Journal of Digital Curation* 3:1, 134–140. ISSN: 1746-8256. URL: http://www.ijdc.net/ijdc/article/view/69 (2009-11-16).

[Hil92] GW Hilton (1992-09). 'A History of Track Gauge'. *Trains*, 23. URL: http://www.trains.com/trn/default.aspx?c=a&id=234 (2010-03-04).

[Hit+05] S Hitchcock et al. (2005-11). 'Preservation for Institutional Repositories: Practical and Invisible'. In: *Proceedings of PV 2005: Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data* (Royal Society of Edinburgh, Edinburgh, UK,). URL: http://eprints.soton.ac.uk/18774/ (2009-11-11).

[Hit+07] S Hitchcock et al. (2007-01-25). *Preservation Metadata for Institutional Repositories: Applying PREMIS*. Draft paper. University of Southampton. URL: http://preserv.eprints.org/papers/presmeta/presmeta-paper.html (2009-09-29).

[HP00] R Heery & M Patel (2000-09). 'Application Profiles: Mixing and Matching Metadata Schemas'. *Ariadne* 25. ISSN: 1361-3200. URL: http://www.ariadne.ac.uk/issue25/app-profiles/ (2008-11-26).

[ISO03a] ISO 14649-1 (2003). *Industrial Automation Systems and Integration – Physical Device Control – Data Model for Computerized Numerical Controllers – Part 1: Overview and Fundamental Principles*. International Organization for Standardization.

[ISO03b] ISO 19115 (2003). *Geographic information – Metadata*. 1st ed. International Organization for Standardization.

[ISO05a] ISO/TS 10303-203 (2005). *Industrial automation systems and integration – Product data representation and exchange – Part 203: Application protocol: Configuration controlled 3D design of mechanical parts and assemblies (modular version)*. International Organization for Standardization.

[ISO05b] ISO/IEC 21000-2 (2005). *Information Technology – Multimedia Framework (MPEG-21) – Part 2: Digital Item Declaration*. 2nd ed. International Organization for Standardization. URL: http://standards.iso.org/ittf/PubliclyAvailableStandards/c041112_ISO_IEC_21000-2_2005(E).zip (2010-09-03).

[ISO06] ISO 10303-224 (2006). *Industrial automation systems and integration – Product data representation and exchange – Part 224: Application protocol: Mechanical product definition for process planning using machining features*. 3rd ed. International Organization for Standardization.

[ISO07] ISO 10303-238 (2007). *Industrial Automation Systems and Integration – Product Data Representation and Exchange – Part 238: Application Protocol: Application Interpreted Model for Computerized Numerical Controllers*. International Organization for Standardization.

[ISO] ISO 10303. *Industrial automation systems and integration – Product data representation and exchange*. Multipart standard. International Organization for Standardization.

[ISO82] ISO 6983-1 (1982). *Numerical Control of Machines – Program Format and Definition of Address Words – Part 1: Data Format for Positioning, Line Motion and Contouring Control Systems*. 1st ed. International Organization for Standardization.

[IV09] P Innocenti & G Vullo (2009). 'Assessing the Preservation of Institutional Repositories with DRAMBORA: Case Studies from the University of Glasgow'. *Bollettino AIB* 49:2, 139–156. ISSN: 1121-1490. URL: http://www.aib.it/aib/boll/2009/0902139.htm (2010-02-19).

[JD08] H James & A Dunning (2008-05-12). *ADS Notes on Writing the AHRC Technical Appendix*. Ed. by K Niven. Archaeology Data Service. URL: http://ads.ahds.ac.uk/project/ahrc/ahrc_guidelines.html (2010-03-25).

[JIS03] Joint Information Systems Committee (2003-06). *JISC Circular 6/03 (Revised): An invitation for expressions of interest to establish a new Digital Curation Centre for research into and support of the curation and preservation of digital data and publications*. URL: http://www.jisc.ac.uk/uploaded_documents/6-03%20Circular.doc (2009-11-11).

[JMK09] S Jones, P McCann & Y Kim (2009-01-22). *A Practical Guide to Implementing the DCC Testbed Methodology*. Version 1.2. Digital Curation Centre. URL: http://www.dcc.ac.uk/sites/default/files/documents/publications/TestbedMethodologyManual_1.2.pdf (2010-03-16).

[Jon09a] S Jones (2009-03-30). *A Report on the Range of Policies Required for and Related to Digital Curation*. Deliverable H1.1. Version 1.2. Digital Curation Centre. URL: http://www.dcc.ac.uk/docs/reports/DCC_Curation_Policies_Report.pdf (2010-06-01).

[Jon09b] S Jones (2009-10). *Data Asset Framework Implementation Guide*. University of Glasgow, Humanities Advanced Technology and Information Institute. URL: http://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf (2010-02-19).

[Jon+09] S Jones et al. (2009-05-26). *Data Audit Framework Methodology*. Version 1.8. University of Glasgow, Humanities Advanced Technology and Information Institute. URL: http://www.data-audit.eu/DAF_Methodology.pdf (2010-02-19).

[Jon10] S Jones (2010-01). *Summary of UK Research Funders' Expectations for the Content of Data Management and Sharing Plans*. Digital Curation Centre. URL: http://www.dcc.ac.uk/sites/default/files/documents/publications/UK%20research%20funder%20expectations%20for%20data%20plan%20coverage.pdf (2010-03-25).

[Kim08] Y Kim (2008-08). *DCC Methodology for Designing and Evaluating Curation and Preservation Experiments*. Version 1.1. Digital Curation Centre. URL: http://www.dcc.ac.uk/sites/default/files/documents/publications/TestBedMethodV1.1.pdf (2010-03-16).

[Kni08a] G Knight (2008-06-10). *Deciding Factors: Issues that Influence Decision-Making on Significant Properties*. Discussion Paper. Version 0.1. JISC. URL: http://www.kcl.ac.uk/content/1/c6/04/55/43/deciding-factors.pdf (2010-01-07).

[Kni08b] G Knight (2008-02-14). *Framework for the Definition of Significant Properties*. InSPECT Project Document. Version 1. JISC. URL: http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf (2010-01-07).

[Kni08c] G Knight (2008-08-21). *Significant Properties Data Dictionary*. InSPECT Project Document. Version 1.0. JISC. URL: http://www.kcl.ac.uk/content/1/c6/04/55/43/sigprop-dictionary.pdf (2009-01-07).

[Kom09] M Komorowski (2009). *A History of Storage Cost*. URL: http://www.mkomo.com/cost-per-gigabyte (2010-03-26).

[Lau08] M Lautenschlager (2008-06). *World Data Center for Climate: Preservation of Earth System Model Data*. Briefing Paper. Digital Preservation Europe. URL: http://www.digitalpreservationeurope.eu/publications/briefs/preservation-of-earth-system-model-data.pdf (2010-03-25).

[Law+09] BN Lawrence et al. (2009). 'Information in Environmental Data Grids'. *Philosophical Transactions of the Royal Society A* 367, 1003–1014. ISSN: 1364-503X. DOI: 10.1098/rsta.2008.0237.

[LM03] P Lord & A Macdonald (2003). *e-Science Curation Report: Data curation for e-Science in the UK: An audit to establish requirements for future curation and provision*. JISC. URL: http://www.jisc.ac.uk/media/documents/programmes/preservation/e-sciencereportfinal.pdf (2009-11-09).

[LN07] H Livingston & D Nastasie (2007-05). 'The Role of Academic Libraries in the Sustainability, Preservation and Access Control of Digital Repositories'. In: *Proceedings of EDUCAUSE Australasia '07* (Melbourne Exhibition and Convention Centre, Melbourne, Australia,). URL: http://www.caudit.edu.au/educauseaustralasia07/authors_papers/Livingston.pdf (2009-11-11).

[LPD05] KH Law, J Peng & P Demian (2005-09-30). *An Assessment of Data Curation Issues for NEES*. Technical Report TR-2005-047. NEES Program. URL: http://wiki.nees.org/download/attachments/689023/TR-2005-047new_Kincho.pdf (2009-11-11).

[LS07] M Lautenschlager & W Stahl (2007-10). 'Long-Term Archiving of Climate Model Data at WDC Climate and DKRZ'. In: *PV 2007: Ensuring the Long Term Preservation and Value Adding to Scientific and Technical Data – Conference Proceedings*. Ed. by E Mikusch & C Reck. German Aerospace Centre (DLR) & German Remote Sensing Data Centre (DFD): Oberpfaffenhofen, Germany. ISBN: 978-3-00-022548-2. URL: http://www.pv2007.dlr.de/Papers/Lautenschlager_LongTermArchivingClimateModelData.pdf (2009-10-21).

[Luc05] A Lucas (2005-11). 'XFDU Packaging Contribution to an Implementation of the OAIS Reference Model'. In: *Proceedings of PV 2005: Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data* (Royal Society of Edinburgh, Edinburgh, UK,). URL: http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/043.pdf (2010-03-09).

[May+95] RJ Mayer et al. (1995-09). *Information Integration for Concurrent Engineering (IICE) IDEF3 Process Description Capture Method Report*. Technical Report KBSI-IICE-90-STR-01-0592-02. Knowledge Based Systems. URL: http://www.idef.com/pdf/Idef3_fn.pdf (2010-03-12).

[Met98] Metadata Ad Hoc Working Group (1998). *Content Standard for Digital Geospatial Metadata*. FGDC-STD-001-1998. Federal Geographic Data Committee. US Geological Survey: Reston, VA. URL: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/ (2010-03-25).

[Mig07] MW Migliaro (2007). 'Standards in Electrotechnology, Telecommunications, and Information Technology'. In: *Standard Handbook for Electrical Engineers*. Ed. by HW Beaty & DG Fink. 15th ed. McGraw-Hill: New York. Chap. 28. ISBN: 978-0-07-144146-9.

[Mil06] P Millington (2006-05). 'Moving Forward with the OpenDOAR Directory'. Presentation given at the 8th International Conference on Current Research Information Systems, Bergen, Norway. URL: http://www.opendoar.org/documents/BergenPresentation20060512Handouts.ppt (2009-10-05).

[MRC05] Medical Research Council (2005-09). *Good Research Practice*. MRC Ethics Series. Medical Research Council: London. URL: http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002415 (2010-04-01).

[MRC08] Medical Research Council (2008-12-19). *MRC Policy on Data Sharing and Preservation*. URL: http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm (2010-01-14).

[MWA06] R McLeod, P Wheatley & P Ayris (2006-04-07). *Lifecycle Information for E-literature: Full Report from the LIFE Project*. Final report. JISC. URL: http://eprints.ucl.ac.uk/1854/ (2008-10-13).

[Nas+08] A Nassehi et al. (2008). 'Toward Interoperable CNC Manufacturing'. *International Journal of Computer Integrated Manufacturing* 21:2, 222–230. ISSN: 0951-192X. DOI: 10.1080/09511920701607899.

[NBJ08] M Nilsson, T Baker & P Johnston (2008-01-14). *Singapore Framework for Dublin Core Application Profiles*. DCMI Recommended Resource. Dublin Core Metadata Initiative. URL: http://dublincore.org/documents/2008/01/14/singapore-framework/ (2009-11-23).

[Neo] *Metadata*. NERC Earth Observation Data Centre. URL: http://www.neodc.rl.ac.uk/popups/faqwindow.php?id=9 (2010-03-25).

[NER02] Natural Environment Research Council (2002-12). *NERC Data Policy Handbook*. Version 2.2. Natural Environment Research Council: Swindon. URL: http://www.nerc.ac.uk/research/sites/data/documents/datahandbook.pdf (2010-01-14).

[Niv08] K Niven (2008-03-26). *Guidelines for Depositors*. Version 1.3. Archaeology Data Service. URL: http://ads.ahds.ac.uk/project/userinfo/deposit.cfm (2010-03-25).

[OCL02] OCLC/RLG Working Group on Preservation Metadata (2002-06). *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of Digital Objects*. Ed. by B Lavoie & R Dale. Online Computer Library Center: Dublin, OH. URL: http://www.oclc.org/research/projects/pmwg/pm_framework.pdf (2009-09-28).

[OMG09] Object Management Group (2009-02-02). *OMG Unified Modeling Language (OMG UML) Superstructure*. Version 2.2. URL: http://www.omg.org/cgi-bin/doc?formal/09-02-02 (2010-03-12).

[Pow+05] A Powell et al. (2005-03-07). *DCMI Abstract Model*. DCMI Recommendation. Dublin Core Metadata Initiative. URL: http://dublincore.org/documents/2005/03/07/abstract-model/ (2009-11-23).

[Pre05] Preservation Metadata: Implementation Strategies (PREMIS) Working Group (2005-05). *Data Dictionary for Preservation Metadata*. Final report. Version 1.0. Online Computer Library Center & Research Libraries Group: Dublin, OH & Mountain View, CA. URL: http://www.oclc.org/research/projects/pmwg/premis-final.pdf (2009-09-28).

[Pre07] C Preston(ed.) (2007-09). *Metadata Encoding and Transmission Standard (METS): Primer and Reference Manual*. Version 1.6. Digital Library Foundation. URL: http://www.loc.gov/standards/mets/METS%20Documentation%20final%20070930%20msw.pdf (2010-03-09).

[PRE08] PREMIS Editorial Committee (2008-03). *PREMIS Data Dictionary for Preservation Metadata*. Version 2.0. Library of Congress: Washington, DC. URL: http://www.loc.gov/standards/premis/v2/premis-2-0.pdf (2009-09-28).

[Rew+10] R Rew et al. (2010-01). *NetCDF Users Guide*. Version 4.1. Unidata Program Center. University Corporation for Atmospheric Research: Boulder, CO. URL: http://www.unidata.ucar.edu/software/netcdf/docs/netcdf (2010-03-25).

[RLG02] RLG/OCLC Working Group on Digital Archive Attributes (2002-05). *Trusted Digital Repositories: Attributes and Responsibilities*. Final Report. Research Libraries Group: Mountain View, CA. URL: http://web.archive.org/web/*/www.rlg.org/en/pdfs/repositories.pdf (2010-02-17).

[RLG05] RLG-NARA Digital Repository Certification Task Force (2005-08). *Audit Checklist for the Certification of Trusted Digital Repositories: Draft for Public Comment*. Ed. by B Ambacher & RL Dale. Research Libraries Group: Mountain View, CA. URL: http://web.archive.org/web/*/www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf (2010-02-17).

[RLG07] RLG-NARA Digital Repository Certification Task Force (2007). *Trustworthy Repositories Audit and Certification: Criteria and Checklist*. Ed. by RL Dale & B Ambacher. Center for Research Libraries & Online Computer Library Center: Chicago, IL & Dublin, OH. URL: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (2010-02-17).

[Rob03] RJ Robbins (2003-10-29). 'Genome Informatics I: Community Databases'. In: *Report of the Invitational DOE Workshop on Genome Informatics, 26–27 April 1993*. Human Genome Project. Baltimore, MD. URL: http://www.ornl.gov/sci/techresources/Human_Genome/publicat/miscpubs/bioinfo/inf_rep2.shtml (2009-12-15).

[Ros+08] C Rosenthal et al. (2008-03-25). *Repository Planning Checklist and Guidance*. Deliverable D3.2. DigitalPreservationEurope. URL: http://www.digitalpreservationeurope.eu/publications/reports/Repository_Planning_Checklist_and_Guidance.pdf (2009-10-05).

[Ros+09] U Rosemann et al. (2009-11-18). *DataCite: Memorandum of Understanding*. URL: http://www.tib-hannover.de/fileadmin/datacite-memorandum.html (2010-03-11).

[SHH10] L Sesink, R van Horik & H Harmsen (2010-02-12). *Data Seal of Approval: Quality Guidelines for Digital Research Data in the Netherlands*. Ed. by H Harmsen et al. Version 2.0. Data Archiving and Networked Services (DANS). ISBN: 978-9490-531-02-7. URL: http://www.datasealofapproval.org/sites/default/files/datakeurmerk_2-0_engels_web.pdf (2010-03-15).

[SN07] C Smythe & B Nielsen(eds.) (2007-03-01). *IMS Content Packaging Specification Primer*. Version 1.2 (Public Draft v2.0). Undergoing standardization as ISO/IEC 12785. IMS Global Learning Consortium. URL: http://www.imsglobal.org/content/packaging/cpv1p2pd2/imscp_primerv1p2pd2.html (2010-03-09).

[Str04] Structural Reform Group (2004-10-06). *DDI Version 3.0 Conceptual Model*. Data Documentation Initiative.

[Tho08] S Thomas (2008-08). *Complex Archive Ingest for Repository Objects (CAIRO) Project*. Final report. Version 0.2. JISC. URL: http://ie-repository.jisc.ac.uk/392/ (2009-09-28).

[Tur03] HW Turner (2003). 'Standards and Certification'. In: *Electrical Engineer's Reference Book*. Ed. by MA Laughton & DJ Warne. 16th ed. Newnes: Oxford & Burlington, MA. Chap. 49. ISBN: 978-0-7506-4637-6. URL: http://www.engineeringvillage.com/controller/servlet/OpenURL?genre=book&isbn=9780750646376.

[Ukda] *Data Lifecycle* (2009-11-02). UK Data Archive. URL: http://www.data-archive.ac.uk/sharing/lifecycle.asp (2010-03-12).

[Ukdb] *Manage and Share Data* (2009-11-03). UK Data Archive. URL: http://www.data-archive.ac.uk/sharing/ (2010-03-24).

[Wel07a] Wellcome Trust (2007-01). *Policy on Data Management and Sharing*. URL: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm (2010-03-19).

[Wel07b] Wellcome Trust (2007-01). *Q & A: Wellcome Trust Policy on Data Management and Sharing*. URL: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/WTX035045.htm (2010-03-19).

[Woo+06] A Woolf et al. (2006). 'Data Integration with the Climate Science Modelling Language'. *Advances in Geosciences* 8:1, 83–90. ISSN: 1680-7340. URL: http://www.adv-geosci.net/8/83/2006/ (2008-12-23).

[WP09] R Williams & G Pryor (2009-11). *Patterns of Information Use and Exchange: Case Studies of Researchers in the Life Sciences*. RIN report. Research Information Network & British Library: London. URL: http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/disciplinary-case-studies-life-sciences (2009-11-27).

[WR07] D Woodyard-Robinson (2007-06-04). *Implementing the PREMIS Data Dictionary: a Survey of Approaches*. Report for the PREMIS Maintenance Activity. Library of Congress. URL: `http://www.loc.gov/standards/premis/implementation-report-woodyard.pdf` (2009-09-29).